FINAL

RECORDS MANAGEMENT STORAGE ARCHITECTURE REPORT
(Work Assignment 008, Deliverable 2-3.1b)

Submitted to:

Ravi Nair
U.S. Environmental Protection Agency
Enterprise Technology Services Division
Research Triangle Park, NC 27711

Prepared by:

SRA International, Inc.
630 Davis Drive, Suite 150
Research Triangle Park, NC 22703

October 1, 1999

EPA IIASC
Contract Number: 68-W-99-038

RECORDS MANAGEMENT STORAGE ARCHITECTURE REPORT

October 1, 1999

EPA IIASC
Contract Number: 68-W-99-038

U.S. Environmental Protection Agency
Enterprise Technology Services Division
Research Triangle Park, NC 27711

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY 1
2 INTRODUCTION 3
2.1 Purpose *3*
2.2 Scope *3*
2.3 Assumptions and Constraints *3*
3 RECORDS 5
3.1 Definition of a Record *5*
3.2 Sources of EPA Records *7*
3.3 How Records at EPA Are Generated *9*
3.4 Importance of Records *10*
3.5 Security of Records *10*
3.6 Longevity of Records *10*
4 KEY ISSUES ON STORAGE AND RETENTION OF RECORDS 12
4.1 Document Inventory *12*
4.2 Backup and Disaster Recovery *12*
4.3 Security *13*
4.4 Recall and Retention Issues *14*
4.5 Conversion Issues *14*
4.6 Electronic Versus Paper Archival *16*
5 EPA'S LONG TERM STORAGE REQUIREMENTS 18
5.1 How EPA Data Are Archived *18*
5.2 Volume of Current EPA Archive *19*
5.3 Projected Volume of Future Annual EPA Archive *20*
6 STORAGE ALTERNATIVES 22
6.1 Storage Technologies *22*
6.2 Storage Media *28*
6.3 Shelf -Life of Storage Media *31*
6.4 Pros And Cons of Storage Media *35*
6.5 Technologies for Image Processing *37*
7 STATUS OF STORAGE SOLUTIONS AT EPA 40
7.1 Mainframe Storage *40*
7.2 LAN Storage *40*
7.3 Distributed Systems Storage *41*
7.4 Image Processing *41*

8 STORAGE MANAGEMENT 42
8.1 Storage Systems Management Technology 42
8.2 Management of Information 44
8.3 Migration of Information 44
8.4 Cataloging, Metadata, and Indexing 45
8.5 Retention 46
8.6 Ready and Reliable Access 47
9 RECORDS STORAGE MANAGEMENT SOLUTIONS 49
9.1 Records Management Software (DoD5015.2 Compliant) 49
9.2 Records Management Software (Other) 51
10 STORAGE SOLUTION CRITERIA AND EVALUATION 53
10.1 Criteria for Storage Systems 53
10.2 Evaluation of Storage Systems Based On Criteria 54
11 RECOMMENDATIONS 57
11.1 Storage Solutions 57
11.2 RMA Software 57
11.3 Architecture Alternatives 58
11.4 Details of Distributed Architecture Alternative 59
11.5 Recommended Future Actions 62

Appendix A. Glossary of Terms 65
Appendix B. Email Records 73
Appendix C. DoD 5015.2 Directive 75
Appendix D. NARA, NRMP and Court Rulings 81
Appendix E. Example Record Schedules 86

1 EXECUTIVE SUMMARY

Electronic records management pressures on the Environmental Protection Agency (EPA) are coming from many different sources both internal and external. Regulatory requirements of the Government Paperwork Elimination Act (GPEA) of 1998 and EPA's Reinventing Environmental Information (REI) initiative encourage electronic forms and reporting. GPEA mandates that electronic reporting capabilities be made fully available to States, facilities, and other stakeholders by 2003. The REI Action Plan requires that all national systems provide at least one electronic reporting mechanism by 2001. Because much of these data are archived at EPA, the need for electronic archival of data is increasing.

Litigation affecting Federal Records Management (see Appendix D) comes from the lawsuit of Public Citizen v. Carlin which declared the National Archives and Records Administration (NARA) General Record Schedule 20 "null and void." NARA's response to "wait and see" received a court reprimand, but an appellate court reversal in August 1999 frees government agencies to make their own decisions on electronic record retention, if paper records are available. However, the sheer proliferation of electronic data is motivation enough, without court intervention, for government agencies to promote electronic archival of record data. In addition, electronic archival brings numerous environmental benefits relating to saving raw materials, chemicals, and energy.

EPA must prepare itself for an increase in electronic storage capacity to accommodate the retention of the ever-increasing number of government records. While NARA has commissioned a fast track group to recommend new standards for electronic archival, EPA must make internal decisions on storage technology and records management because a large quantity of data will be retained in EPA systems on average between 4 to 6 years.

This paper addresses five areas of the storage problem: 1) what storage media is best suited for retention of data, 2) what storage system is best suited for digital records processing, 3) what applications should be investigated to manage the records retention process, 4) in what format digital data should be archived, and 5) what is the optimal records management and storage architecture.

Selecting the best storage solution depends on the media and the storage system and how these satisfy the storage purpose. The criteria considered in this study were access time, capacity, cost, availability, portability, scalability, security, and longevity.

This study concludes that optical media currently provide the best durability, longevity, security, and scalability of the media available for electronic records. In the higher densities available with digital versatile disk (DVD), optical media is also cost competitive with magnetic tape. Although magnetic tape media has the lowest cost per megabyte, it is a secondary choice because of drawbacks associated with its lack of durability, longevity, and random file access.

For external archival of EPA data in digital form, DVD is the study's recommended optical medium; however, NARA is currently evaluating optical media and a decision on which type to use should comply with their recommendations.

This study concludes that an optimal approach to high capacity storage systems is a jukebox design that can handle from 200 to 600 or more optical disks. A jukebox storage system should be supplemented with hard disc caching for both access time and cataloging considerations. A properly cached storage system will provide good storage access and availability. For faster on-line storage of active records with transparent backup, a Storage Area Network (SAN) with Fibre Channel architecture is desirable.

EPA also needs to acquire an integrated software product to manage the storage, retention, and retrieval of electronic records. Without a Records Management Application (RMA), the task of tracking and retrieving electronic records during their lifecycle looks daunting, at best. Selection of an RMA should not necessarily be tied to selection of a storage device or storage system. RMAs should be evaluated based on their data management, ease-of-use, and workflow process capabilities. As such, an RMA would preferably be storage-independent. Electronic archival requires a Records Management Application to coordinate the creation of metadata and the storage, retention, retrieval, and archival of data in a user-friendly fashion. Selecting an RMA will require an in-depth requirements analysis and an assessment of integration with EPA's information technology (IT) architecture.

Part of the process of archiving digital data is converting existing digital data to a digital archival format, and converting data in paper or microfilm form to that same digital archival format. A consistent well-supported format becomes critical to information accessibility and usability. This common digital format would actually be a digital "image" of the document. This study recommends the use of Tagged Image File Format (TIFF) and Portable Document Format (PDF). A recent study by the Department of Defense listed TIFF and PDF as the two most common image formats in use by the government today.

Finally, the study recommends a *distributed* records management and storage architecture. Such an approach makes each regional office responsible for the proper administration, maintenance and support of their records program. Due to the large volume of records created in the Agency, and the often highly sensitive nature of their content, it is most practical to keep the information as close to the record custodian as possible.

2 INTRODUCTION

The Environmental Protection Agency (EPA) is involved in developing a records management process to comply with internal and external requirements for the retention and retrieval of information pertinent to the mission and operation of the Agency, and the heritage requirements of the federal government.

The SRA Team is providing this analysis paper for the Office of Policy (OP) under the framework of IIASC Work Assignment 8, Task 2: Technical Consulting and Analysis for Key EPA Technologies.

2.1 Purpose

The purpose of this study is to investigate EPA's records storage and management requirements as they pertain to electronic records and storage technology. This study will help EPA analyze retention and retrieval issues in the areas of records storage and management, and this study will provide an analysis of digital storage options available to EPA.

2.2 Scope

The electronic records management (ERM) challenge is being addressed by the National Archive and Records Administration (NARA) and by EPA's National Records Management Program (NRMP). This study addresses EPA's requirements for Electronic Records Management (ERM) and digital storage from a high-level technology viewpoint. It is understood that NRMP formulates guidance for the Agency and any recommendations made in this paper are subject to this guidance.

2.3 Assumptions and Constraints

- This study will examine the storage of archived records only in the context of the ability to view those records. It does not address re-usability issues of the records being archived.

- For the purposes of this document, data in its working form is not considered a record. However, any final report (including data) could be considered a record.

- Security of records being transferred across network connections is outside the scope of this document, and is governed by applicable Agency IT security policies and procedures.

- Backup and disaster recovery of records are discussed only as they pertain to archived records stored off-line at EPA. It is assumed that the disaster recovery plan governing all general support systems will cover records and records metadata stored elsewhere within the Agency.[1]

[1] General Support (GS) systems refers to the systems in place that constitute EPA's information technolo infrastructure. The GS systems include the mainframe, desktops, scientific workstations, supercompute networks and local area networks, network servers, and Internet/Intranet servers. Each of these compone their own plans for security and disaster recovery. A record management system will rely on many of th components to function, and the security and backup of records in these components will be governed b in place for those components.

3 RECORDS

This section provides background information on what a "record" is, what it isn't, types of records, their formats, security, and how records are generated at EPA.

3.1 Definition of a Record

In the most general terms, Gartner Group defines a record as "any information-bearing media generated or received by an organization."[2] As more strictly used in the government, a "record" is the term used for information that should be archived. The term "record" does not imply in what form or for how long information will be stored.

EPA's Directive 2160 Records Management Manual defines a "record" as follows:

"Recorded information of continuing administrative, fiscal, legal, historical, or informational value, including published materials, papers, maps, photographs, microfilm, audiovisual, machine-readable materials (ADP tapes/disks) or other documentary material, regardless of physical form or characteristics, made or received by the Agency that evidences organization, functions, policies, decisions, procedures, operations, or other activities of the Government."[3]

Furthermore, the manual describes the following specific types of records:

- **Classified Records.** Records designated as "Top Secret," "Secret," or "Confidential" which are restricted to processing or use by cleared individuals, and require special protection.

- **Current Records.** Records or files presently in the physical custody or organizational units, the maintenance of which is required for the conduct of current work.

- **Non-record Material.** Material of this nature includes blank forms, library materials, and working papers of fleeting value such as drafts, worksheets, informal notes, slips, etc.

- **Official Record File.** The complete documentation including all background material resulting from specific transactions, operations, or processes which are accumulated and maintained in files equipment. They may include any media such as film, microfilm, cards, papers, and magnetic tapes and disks.

---

[2] Records Management: Eight Steps to Control Office Chaos, M. Gilbert, Gartner Interactive, Gartner Group, © August 1998.

[3] http://www.epa.gov/irmpoli8/recmgmt/chaptr1.txt.html

- **Permanent Records.** Records of continuing value which are considered to be so valuable or unique in documenting the history of the Agency or for informational content that they should be preserved "forever" as part of the National Archives of the United States.

- **Record Copy.** The official file copy, either originated or received by any element of EPA, which documents an action relating specifically to the functions assigned by operating elements in the Agency.

- **Semi-active and Inactive Records.** Records worthy of preservation and which have long term permanent value will be retired from expensive office space and equipment to the area Federal Records Center for storing, servicing, and ultimate disposition in accordance with EPA records control schedules.

- **Temporary Records.** Records created incidental to performance of the mission. They are "operational," "support," and "service" type records which are considered to be of temporary value to the Agency and will be destroyed at some time.

3.1.1 What Is a Record?

The NRMP has defined a "record" according to a series of questions and answers. Official Records[4] are any final product related to EPA's administrative, enforcement, or regulatory policies and activities. Examples of final products include:

- Decision papers
- Memoranda
- Letters
- Films
- Data files
- Models
- Reports
- Publications

These supporting materials are sufficient to document and/or explain the document trail/decision-making process for administrative, legal, final, programmatic, and historical purposes. They may include drafts, annotations, reports, raw data, meeting minutes, and telephone logs. Official records may be originals or copies of original records.

[4] http://www.epa.gov/records/whatis/yes.htm

3.1.2 What Isn't a Record?

According to NRMP's question and answer guidelines, non-record[5] information is the following:

- Materials that do not contribute to an understanding of EPA operations or decision-making processes.
- Materials that have no substantial programmatic value.
- Exact (extra) copies of official record documents retained elsewhere that serve as:
  - Convenience copies kept solely for ease of access and reference.
  - Information/reference copies of records sent to individuals or offices interested in, but not acting on, a matter.
  - Technical reference documents needed for general information, but not properly part of the office's records.

According to the NRMP website, "Each employee is responsible for determining if the documentary materials they have are official Agency records, non-records, or personal papers."[6]

3.2 Sources of EPA Records

EPA generates records from many sources in operational and administrative areas. The operational sources include the scientific collection and analysis of data, regulatory requirements and directives, and the published and unpublished results of studies. Administrative sources include email, human resource information, financial data, policies and procedures information, and internal directives.

The Gartner Group states that by the end of 1999, approximately 75% of regularly used document information will be stored in digital form.[7] It is reasonable to assume based on the knowledge of the Agency that this percentage is also a valid statistic for EPA document information. Gartner also states that email integration with document management will be the focus of records management vendors in the new century. With this in mind, it is clear that the majority of EPA records are, or soon will be, in electronic (digital) form. The past practice of warehousing boxes of paper documents will no longer be a viable approach to handling the records EPA is generating.

---

5 http://www.epa.gov/records/whatis/no-non.htm

6 http://www.epa.gov/records/whatis/index.htm

7 Records Management: Eight Steps to Control Office Chaos, M. Gilbert, Gartner Interactive, Gartner Gr
August 1998.

Sources of electronic records at EPA include, but are not limited to, the following:

- **Internet Websites**. The EPA Internet website is a huge collection (currently 246,000 files) of electronic records maintained and made available for the public. Currently, EPA has a draft website archival schedule (INFO-095) that provides useful guidance, but it has not yet been approved by NARA. More and more of EPA's documents and reports are being posted on the Internet or Intranet for easy access, as well as kept in more traditional project-centric electronic directories.[8]

- **Internet/Intranet Metadata**. Information about EPA's website data is in the process of being cataloged in Oracle databases, and is not currently printed on paper for storage.

- **Email**. This includes archival files from GroupWise and Notes Mail systems, and non-archival information as well, and is printed on paper only sparingly.

- **Notes Databases**. This includes numerous discussion forums and decision databases for technology choices and other non-technology topics, and is rarely printed on paper for storage.

- **Documents**. This includes reports, assessments, plans, guides, workbooks, project findings, study findings, and other formal papers that have been wordprocessed and are available in electronic and paper format.

- **Databases**. This includes Oracle, Personal Oracle, Approach, DB2 and other sets of electronic information maintained for fast access. The information ranges from scientific to administrative and is sometimes printed on paper for storage.

- **Graphics**. This includes maps, map objects, environmental visualizations, diagrams, layouts, drawings, etc., that are primarily in electronic format and are often printed to paper format.

- **Data Files**. This includes working data that range over all EPA functions and are most often stored electronically in flat files, spreadsheets, text files, and array files. Depending on the function, these data may be printed on paper for storage, but are most often represented only in electronic format.

3.3 How Records at EPA Are Generated

Material classified by EPA as a "record" comes from two fundamental sources:

- External sources (e.g., States, tribes, other federal agencies, the regulated community, environmental partners, and the general public).

[8] Records Management for EPA Web Sites, 1999 EPA Records Management Conference, Carolyn Offutt 1999.

- Internal sources (e.g., Program offices, Headquarters, Regions, and labs).

The information generated by internal EPA sources comes from a variety of software. This information is in a wide variety of formats, even though most of it is in electronic form. The Agency standard COTS applications that account for much of these data are listed below:

- Lotus Notes Mail
- GroupWise Mail
- Lotus 123
- Lotus Freelance
- Lotus WordPro
- Lotus Approach
- Corel WordPerfect
- Adobe Acrobat

In addition, large amounts of information are generated, stored, and managed by Agency-developed applications systems. Some of these are:

- Aerometric Information Retrieval System (AIRS)
- Comprehensive Emergency Response and Liability System (CERCLIS)
- Integrated Contract Management System (ICMS)
- Integrated Data for Enforcement Analysis (IDEA)
- Integrated Financial Management System (IFMS)
- Integrated Risk Information System (IRIS)
- Permit Compliance System (PCS)
- Pesticide Product Information System (PPIS)
- Resource Conservation and Recovery Information System (RCRIS)
- Safe Drinking Water Information System (SDWIS)
- Storage and Retrieval of Water Quality Information (STORET)
- Superfund Cost Recovery Image Processing System (SCRIPS)
- Toxic Chemical Release Inventory System (TRIS)

Information classified as records by EPA that come from external sources are generated from an even broader set of software and systems (e.g., Agency non-standards products such as Microsoft Word and Excel, Electronic Data Interchange transactions, proprietary state reporting systems). EPA often has no control over the software used to generate information that is submitted by external entities, yet much of this information must be managed as official records by the Agency. In addition to electronic sources, a great deal of information to be classified as records is still submitted in paper form.

Regardless of the source (internal or external, electronic or paper, Agency standard or non-standard), the process adopted for electronic storage and retrieval should neither interfere with nor restrict the present approved policies and processes currently in place for records generation. The ultimate solution selected must allow for the capture, translation, and secure storage of record items generated

by all designated sources with complete fidelity, while providing a fast and reliable retrieval mechanism. This issue is addressed further in section 4 of this paper.

3.4 Importance of Records

Some records are more important to the Agency than others; that is, they are more critical to the functioning of the Agency and its ability to support the environmental mission (e.g., studies, test data, regulations, directives, legal casework, etc.). Other critical data involve the administrative functioning of the Agency (e.g., payroll, human resource records, procedures, guidelines, etc.).

NRMP assists offices to determine how different types of records should be stored. They provide assistance with designing record schedules to properly handle important information.

3.5 Security of Records

Each office is responsible for designating in their record schedules the security requirements and confidentiality of the data being retained by EPA. For instance, some schedules require destruction of confidential business information (CBI) data rather than archival, and when the destruction of data is handled by contractors, they require the presence of EPA as witnesses. Once data have been archived to Federal Records Centers (FRCs), the security policies and procedures of the FRCs and NARA prevail.

3.6 Longevity of Records

Each office at EPA is responsible for designating how long to retain data in their record schedules. This pertains to data kept in the offices as well as to data sent to the FRCs for archival. Record schedules must clearly specify how long data are kept in each office and when they are to be destroyed or archived, and if archived, how long the FRCs should retain the data. In general, records should be kept for as long as business, regulatory, or historical purposes require.

There is no uniform archive schedule for environmental data. Retention of environmental data at EPA is governed by numerous Record Control Schedules put in place by the respective programs and approved by NARA. See Section 5 and Appendix E for examples of these schedules.

4 KEY ISSUES ON STORAGE AND RETENTION OF RECORDS

4.1 Document Inventory

One of the first and most critical steps in creating and maintaining a Records Management system is to create a document inventory. A document inventory can be used to identify the flow of documents, the retrieval and access patterns, and many other critical document attributes. The components that make up this inventory include Document Purpose, Document Creator, File Format (what application/version the document was created in), File Volume (number of chapters and pages), Security (does the document contain any confidential business information), Reporting (usage and viewing statistics), and Retention/Purging requirements. Within the Agency, this type of document inventory is often referred to as a "metadata record," which describes the document's characteristics at the time of its creation. Most Record Management Systems maintain this inventory as part of their application.

The document inventory process will require a substantial amount of effort, but will provide an invaluable tool for managing and retrieving records.

4.2 Backup and Disaster Recovery

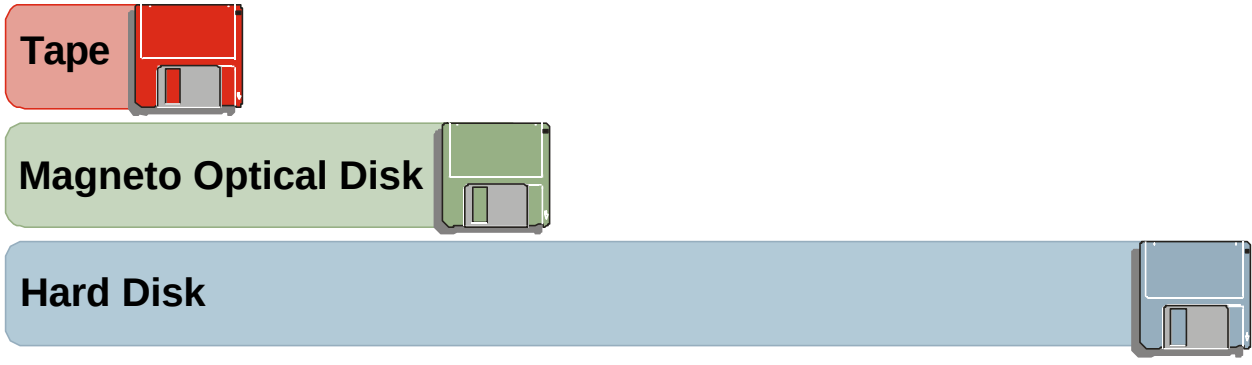
There are three primary causes of data loss that should be of concern to the Agency: 1) System failure/malfunction, 2) File Corruption (through malicious sabotage such as a virus or theft, willful alteration, or un-intentional such as media corruption or accidental erasure), and 3) Natural Disaster.

The backup/disaster recovery policies and procedures governing the General Support (GS) systems in use at the Agency will provide protection from the above threats as long as documents remain on the GS systems. As documents are moved off-line, and become records for long-term storage, this protection is lost. To either eliminate, or minimize the impact of these cases of data loss, removable data storage is imperative for archival of records. As these records are migrated to removable media, a primary consideration for EPA will be whether or not to maintain backups of these archived records. In essence, this backup process would mean creating a duplicate copy of the record and storing it in an alternate facility.

Tape has been, and will be for the foreseeable future, the lowest cost storage medium. If you look at the cost of storage today, the *incremental* cost per megabyte of hard disk is $0.0400, Magneto Optical Disk is $0.0150, and tape is $0.0030. Tape storage is approximately 1/10 the cost of hard disk storage.[9] This storage comparison is graphically represented in Figure 4-1.

[9] © 1998 Exabyte Corporation. All rights reserved. Incremental cost does not include cost of the storage the cost of the media.

Figure 4-1. Comparative Incremental Costs of Removable Storage Media



Removable media offers many advantages. Removable media allows the administrator to revert back to an earlier copy of any file in the case of viruses, sabotage, file or database corruption, human error, or new problems such as Y2K. Furthermore, with a good media rotation schedule that includes off-site storage, it protects against natural disaster, theft, and sabotage. Since tape is the most economical media, it makes sense to consider using tape as a backup to whatever media is selected for long-term records storage at EPA.

4.3 Security

Security as it relates to a records management system involves the structure of permissions that allow for creating, reading, writing, updating, and deleting of records. While these records must be made available to the public, assurance is needed that these records will not be altered or compromised in any fashion. This is one area where the use of some optical media has a clear advantage over magnetic tape. Utilizing recordable, write-once optical devices ensures that records are always in the same format they were originally stored and that data has not been altered or compromised in any way, since the optical media cannot be re-recorded.

A second concern in the area of securing records is the issue of confidential business information. This data must be compartmentalized, so all CBI data can be stored in a separate area with access only by authorized individuals. CBI data may also require different retention time and disposal methods than other EPA data, or may require a different medium such as one that cannot be accidentally erased.

4.4 Recall and Retention Issues

The key to successful record recall lies in the creation of appropriate indexing and cataloging structures. The catalog structure would consist of an inventory for each individual record. The ability to search for records is dependent on developing indexes and catalogs that reflect the ways that users prefer to search for documents. This requires up-front research and planning, but will ensure the back-end success of the records management system. This planning process needs to include such items as

defining the set of attributes that will be tracked for each record, and creating a document index data dictionary that includes all relevant attribute information.

Part of the purpose for the catalog will be to establish the retention period for that record. EPA will need to decide if the authors of the records will determine this based on Agency guidelines, or if the Agency will automatically assign retention periods based on the type of record as classified by the authors.

Procedures must be put in place to migrate records during the retention periods. As retention periods expire records must be deleted and catalogs updated.

A good records management software application is critical to manage both the recall and retention processes. These applications are discussed further in Section 9 of this paper.

4.5 Conversion Issues

Part of the process of archiving digital data is converting existing digital data to the digital archival format, and converting data in paper or microfilm form to that same digital archival format. A consistent well-supported format becomes critical to information accessibility and usability.

As the first step in the conversion process, a digital archival format must be chosen. While it is possible to store digital data in its native format (i.e., a WordPerfect 8 file, or a Freelance Graphics file) this creates problems with discontinued software, and possibly problems as newer versions of software are published. Historically, newer software versions have been backwards compatible. However, there are no guarantees this trend will continue indefinitely. This highlights the need for a common digital format for archival. This common digital format would actually be a digital "image" of the document. A recent study by the Department of Defense[10] listed Tagged Image File Format (TIFF) and Portable Document Format (PDF) as the two most common image formats in use by the government today.

TIFF was created by Aldus Corporation and Microsoft in the late 1980s to promote the interchange of digital image data. Recently, the Adobe Corporation has merged with Aldus, and has acquired the rights to TIFF. The format creates an exact replica of what the document would look like if it were printed. TIFF is supported by nearly every scanner manufacturer, which would allow easy conversion of printed documents to electronic format. Documents currently stored electronically could be easily converted to TIFF. Many major applications support the conversion to TIFF natively, and software is readily available to perform the conversion if the application itself does not support it.

---

[10] *Electronic Digital Imaging Standards for Archiving Records*, Final Report, June 1, 1999 written for t Department of Defense under Contract GS35F4863G by Susanne H. MacTavish and Michael R. Pickard Lockheed Martin Technical Services, Falls Church, Virginia.

TIFF is primarily a gray scale image format, though extensions of the format provide for color. Since color is becoming more prevalent in printed documents and very prevalent in electronic documents, the selection of an imaging format should include capability for faithful color imaging. TIFF with CCITT (Consultative Committee on International Telephone and Telegraphy) Fax 4 Compression is ideally suited for black and white text documents. This format provides a high level of detail (600 dpi), combined with a small file size (less than 100 kilobytes for 5"x8" text page). It may be used as an archival file format. TIFF with LZW (Lempel-Ziv Welch) Compression is a 24-bit, lossless (no information lost) compression format commonly used by Adobe Photoshop. This variation of TIFF can be used to store color images, and can be used as an archival file format. With lossless compression, the picture quality of the compressed file is exactly the same as the original uncompressed file.[11]

Adobe Corporation's PDF is a completely device-independent page description language. This has made for wide acceptance of PDF as the standard for publishing documents on the Web. PDF documents are an exact replica of the original document, but are smaller in file size than the original document format, thereby reducing the amount of storage space required. PDF is a proprietary format; however, Adobe has made the PDF standards available to many other vendors who have created products to read PDF files. Adobe Acrobat Portable Document Format provides a convenient way to view and print images at high resolution and in color as well.

Records that will be converted must be cataloged and indexed. If these steps are not taken, the records themselves will be preserved, but access to any particular record would be very difficult, if not impossible.

Any storage solution should consider the hardware and software to read the physical media and the data format used, and that these might have to be refreshed as technology changes. Using TIFF or PDF does not preclude such considerations.

4.6 Electronic Versus Paper Archival

Currently, EPA archives most of its records in paper format. This is not uncharacteristic of federal agencies because NARA accepts records electronically in very limited and archaic formats. However, regulatory requirements of the Government Paperwork Elimination Act (GPEA) of 1998 and EPA's Reinventing Environmental Information (REI) initiative encourage electronic forms and reporting. GPEA mandates that electronic reporting capabilities be made fully available to States, facilities, and other stakeholders by 2003. The REI Action Plan requires that all national systems provide at least one electronic reporting mechanism by 2001. Because much of these data are archived at EPA, the need for electronic archival of data is increasing.

[11] Technical Recommendations for Digital Imaging Projects, http://www.columbia.edu/acis/dl/imagespec.

The desire of federal agencies to have new guidelines for electronic records has led to formation this spring of the Federal Interagency Records Management Council (FIRM).[12] There is speculation that FIRM's creation is meant to apply organized pressure on NARA to streamline its processes for electronic records. In the meantime, NARA is also under pressure from the DoD to arrive at an electronic records standard because DoD has millions of records they would like to archive, but they are unwilling to archive these records in paper format.

In addition to regulatory pressures and the trend in government agencies, there are numerous benefits to electronic records archival, some of which are environmental (and therefore warrant additional attention by EPA). The benefits are listed below:

1. Storing records electronically significantly reduces paper and cardboard consumption, and saves the destruction of trees and lowers water and air pollution caused by the production of paper.

2. Electronic records require less physical storage space both within EPA and at NARA. Data densities are much higher for electronic media than for paper. This can reduce cost of storage.

3. Electronic records are easier to locate and retrieve than records kept in paper boxes, particularly when searching for information that is stored in different records.

4. Retrieving electronic records can be much faster than retrieving paper records.

5. Once retrieved, electronic records can be delivered electronically in paperless form, while paper records must be copied or scanned before delivery. Electronic delivery reduces paper usage and cost.

6. Data in electronic record form are more usable than in paper form. For example, text in electronic form can be copied and pasted, while in paper form it must be retyped or scanned with OCR software.

7. Electronic records are easier to transport and handle, particularly if electronic transmission is used. Energy costs and the pollution associated with moving paper are avoided.

Electronic records archival has a few disadvantages. They include:

1. Electronic media are generally not as durable as paper, and the durability (> 20 years) for electronic media is unproven, while paper lifetimes are well established at over 500 years.

[12] http://www.fcw.com/ref/hottopics/records//background/fcw-sprehe-5-31-1999.html

2. Electronic media are more susceptible to data loss from humidity and rough handling, and generally must be kept in well conditioned environments.

3. The technology that reads electronic records may become obsolete or not function correctly without expert personnel, while reading paper requires no special technology.

4. Some electronic media are susceptible to damage from magnetic fields or static electricity. Paper is not adversely affected by these.

In general the advantages of electronic media outweigh the disadvantages for most types of data. Another reason for using electronic record storage is the sheer volume of data that are being earmarked for archival. Most federal Agencies see an electronic storage method as the only effective means of dealing with the profusion of data caused by the increasing usage of information processing systems.

5 EPA'S LONG TERM STORAGE REQUIREMENTS

EPA's long term storage requirements will be dictated in part by the volume of data that the Agency has to place in archival storage each year. The greater the volume of data, the more pressing the need for an electronic records management system and digital storage medium.

5.1 How EPA Data Are Archived

EPA maintains through the EPA National Records Management Program (NRMP), a series of Record Schedules that define how and when to dispose of data. Disposal often requires archiving to a Federal Records Center (FRC) then to NARA. Appendix E, Example 1 describes this scenario. Disposal can also mean destruction of the data as shown in Appendix E, Example 2.

Each office is responsible for implementing an appropriate Record Schedule. NRMP lists the approved schedules on the Web at http://www.epa.gov/records/policy/schedule/sch-titl.htm. As of August 31, 1999 there were 609 active and 60 draft Record Schedules at EPA.[13] The archival effort is very distributed within the Agency. Offices usually send their archival data to regional FRCs according to their schedules for the data they handle. Records in the FRC are eventually destroyed or sent to NARA.

Table 5-1 below provides a look at EPA archival. As can be seen from the Disposition Instructions, there is great variety in the retention periods for data held by EPA.

Table 5-1. Sample U.S. EPA Records Control Schedule Disposition Instructions

Series No.	Series Title	Medium	Disposition Instructions
516	Federal Facilities Referrals File	Paper	Keep inactive materials in office 1 year, then retire to FRC. Destroy when 5 years old.
413-C	Federal Reporting Data System (FRDS)	Electronic	Transfer ASCII or EBCDIC flat file of data on new or recertified one-half inch, 9 track tape or cartridge annually to the National Archives along with tape layout, block size, and other tape specifications.
012-a	Federal Agency Hazardous Waste Compliance Site Files	Paper	Keep in office at least 1 year after file break, then retire to FRC. Transfer to the National Archives 20 years after file break.
012-d	Federal Agency Hazardous Waste Compliance Site Files	Microform	Keep in office until issuance of ROD for the site or operable unit and appropriate milestones thereafter, then retire one silver and one diazo to FRC. Transfer to the National Archives 20 years after file break.

[13] http://www.epa.gov/ngispgm3/nrmp/news/saa/saa.htm

Series No.	Series Title	Medium	Disposition Instructions
012-e	Federal Agency Hazardous Waste Compliance Site Files	Optical or Magnetic media	Delete records when no longer needed.
258-a	Final Deliverables and Reports - Programmatic or mission related	Paper	(Non-Superfund) Keep inactive materials in office at least 1 year after file break, then retire to FRC. Transfer to the National Archives 20 years after file break.
258-a	Final Deliverables and Reports - Programmatic or mission related	Microform	(Non-Superfund) Retire one silver and one diazo to the FRC along with finding aids and indexes. Transfer to the National Archives 20 years after the file break. Retain one or more sets for office use.
258-b	Final Deliverables and Reports -Non programmatic or administrative	Paper Microform Magnetic Optical	Keep inactive materials in office at least 1 year after file break, then retire to FRC. Destroy 7 years after file break.
527	Grants File - Congressional Data Transfer	Paper	Keep inactive materials in office up to 5 years after file break, then retire to FRC. Transfer to the National Archives when 20 years old.
231	Regional Presiding Officer Case Files	Paper	Keep inactive files in office until all required action items have been completed, then retire to FRC. Transfer to National Archives 10 years after file break.
155-a	Reports to Congress and/or the President	Paper	Keep inactive materials in office up to 5 years after file break, then retire to FRC. Transfer to the National Archives in 5 year blocks 20 years after file break.
305	Requests for ADP Hardware and Software	Paper	Keep inactive materials in office for 3 years after file break, then destroy.
227	Toxic Substances Facility Files (Regional)	Paper Electronic	Keep inactive files in office up to 5 years after file break, then retire to FRC. Destroy 10 years after file break.

As required by 36 CFR 1228.58, records with special restrictions such as CBI or those exempted from disclosure by statute such as the Privacy Act must be shredded or otherwise definitively destroyed with witnessed disposal for records destroyed by contractors.

5.2 Volume of Current EPA Archive

Most of EPA's archival content is in paper format. In 1998, EPA paid NARA for cumulative archives of 142,098 cubic feet of information, at a cost of $3.28 per cubic foot, which includes storage and service charges.

To translate cubic feet of paper information into a rough estimate of a quantity of digital data, the following equivalencies were used:

1 cubic foot of paper = approximately 4,000 sheets of paper

1 sheet of paper = approximately 24,763 bytes of data[14]
50% of sheets = single-sided printing
50% of sheets = double-sided printing

Using the above equivalencies, and if we assume an all-paper archive, the rough estimate in electronic bytes of EPA's data archive stored by NARA is 142,098 cubic feet × 4,000 paper sheets/cubic foot × 24,763 bytes/sheet × 1.5 (factor for double-sided printing) = $2.1112636644e^{+13}$ bytes of data or 20,134,579 Megabytes or 19,663 Gigabytes.

In very rough terms, EPA has about 19,663 Gigabytes or 19.2 Terabytes of data held by NARA. A study is needed of the 600 plus Record Control Schedules to determine what percentage of data is sent to NARA and what percentage of data is retained by EPA and not sent to NARA. This would provide for a better estimate of the volume of data that EPA is retaining internally, and is necessary to properly size a storage system solution acquired for archival of data.

5.3 Projected Volume of Future Annual EPA Archive

The following historical data were obtained for records that have been archived by EPA Headquarters office to the Washington National Records Center (WNRC) in Suitland, MD from Fiscal Year (FY) 1989 to FY 1998.

Table 5-1. Retirement of Records from EPA Headquarter Offices By Year

Year	Cubic Feet
1989	1,991
1990	2,015
1991	966
1992	3,039
1993	4,692
1994	5,382
1995	15,667
1996	6,011
1997	4,516
1998	2,866

The above data are not necessarily predictive of the Agency as a whole; however, these data can be used as an example for calculating EPA's annual archival needs, as follows. Eliminating the highest

[14] Printed paper with just text contains about 2,800 characters, but many documents contain tables and figu significantly increase the character count. Also, wordprocessor formats contain many additional format characters. The number 24,763 was obtained empirically by averaging pages for several documents with used for electronic storage of those documents in the WordPerfect 8 file format. (Total of 230 pages u bytes).

and the lowest years from the HQ data (non-characteristic years), the average yearly archival for HQ is calculated to be 3,814 cubic feet (25,715 / 8). Since HQ archival represents approximately one third of EPA's total archive at NARA, we will assume EPA's total transmittals to NARA to be three times the HQ annual archival or 11,442 cubic feet per year.

Using the equivalencies from the previous section to convert cubic feet to bytes, we can calculate a rough estimate of annual archival in bytes as follows: 11,442 cubic feet × 4,000 paper sheets/cubic foot × 24,763 bytes/sheet × 1.5 (factor for double-sided printing) = 1,700,029,476,000 bytes or 1,583 Gigabytes.[15]

These numbers represent rough estimates and are based only on the equivalency of electronic documents data (stored in their native application format) versus printed paper. Additional analysis (beyond the scope of this paper) of archival data, archival formats, and record schedules is needed to determine more accurately EPA's current and future archival volumes.

[15] The conversion here of bytes to Kilobytes to Megabytes uses 1024 as the division factor instead of 100 uncertainty in professional computer circles as to whether 1024 or 1000 should be used as both factors the industry by different groups. There is an international standards effort underway to clarify the conve to higher multiples.

6 STORAGE ALTERNATIVES

There is currently tremendous focus in the IT industry on storage technology, storage management, and information management. Therefore, the range of options in both storage technology and approaches to storage management is very large.[16] With so much to consider, in-depth analysis is needed to match a solution to specific requirements. The purpose of this section is to provide an overview of the mainstream storage technologies available today.

6.1 Storage Technologies

Information storage is divided into **on-line**, **near-line**[17], and **off-line** types. A storage solution may use one, two, or all of these types. Because the cost per megabyte is different for each type, EPA should carefully evaluate its data storage needs before selecting a solution.

A well-conceived storage management system optimizes all three data storage types by ensuring that each contains data available at access speeds that correlate to the data's value and frequency of use in the organization. Combining these three storage types into one system is referred to as Hierarchical Storage Management (HSM).

There are three fundamental factors in selecting a particular storage media: **capacity**, **access time**, and **cost**. To paraphrase a computer industry adage, "you can have storage systems that are high capacity, fast access, or low cost; pick any two." This illustrates the trade-offs involved. Availability is an additional factor that affects the type of storage that is selected.

The selection of storage devices is also driven by the nature of information, the volume of information, and the life cycle of information. For example, Confidential Business Information (CBI) or legal data may benefit from using a non-erasable media for storage. Volume of information affects the economics of the storage solution and thus determines the choice of one media over another. Also, "at certain stages in its life cycle, information is subject to different usage patterns."[18] For example, a mortgage application will be accessed repeatedly prior to and immediately following the closing; however, six months after closing, access will be very infrequent.

---

[16] A web search for "storage management" resulted in hundreds of companies and thousands of links for st management software, storage devices, and storage services.

[17] A storage unit that is connected to a computer system through an automatic storage interface system. E StorageTek silos or CD Jukeboxes.

[18] Storage Management Technology Guide, Copyright 1999, The Applied Technologies Group, http://www.techguide.com . Provided by OTG Software, http://www.otg.com

Storage devices can be classified into three major groups: **semiconductor memory devices**, **rotational devices**, and **linear devices**. Some of the devices in each group are:

- semiconductor memory: ROM, PROM, RAM, CCD, and Bubble Memory
- rotational devices: magnetic disk, optical disks (CD/DVD-ROM, CD-R, etc.), RAID
- linear devices: magnetic tape (9-track reel, DAT, DLT, RAIT, etc.)

In addition to the basic storage devices, three technologies that are currently prominent are RAID (Redundant Arrays of Independent Disks), RAIT (Redundant Arrays of Independent Tape), SAN (Storage Area Networks) and Fibre Channel. RAID and RAIT improves the fault tolerance and speed of magnetic storage and Fibre Channel improves the speed and geography of communicating with storage devices. SAN combines network infrastructure and storage to provide users with a "storage utility".

6.1.1 RAID

RAID is an acronym for Redundant Arrays of Independent Disks. Fault tolerance and fast access times are two of the desirable attributes of RAID technology. Invented at Berkeley in 1987, RAID is one of the computer industry's youngest technologies. The first commercially available RAID disk arrays were introduced in 1995. Since then, RAID has become one of the fastest growing technologies in an already exploding industry, with estimated 1996 worldwide sales reaching $12.3 billion.[19]

RAID is referred to in levels of implementation from "0 " to "7." The most popular implementations are RAID 1 and 5. RAID 1 simply consists of two disks of equal size with one disk being a complete copy of the other. If disk problems occur on one disk, the information desired should be available on the other disk. RAID 5 is much more sophisticated as it employs "striping," or a way to divide a file record across more than one disk with an additional disk for error correction codes (ECC). If a disk fails, only part of the file record is lost, and that part can be re-created using the ECC information. If the ECC disk fails, it can be re-created from the data disks containing the record.

The redundancy feature of RAID systems provides a layer of protection against hard disk failure and provides quick recovery from such failure. For this reason and the high throughput, RAID systems are often considered for reliable on-line access to large quantities of data.

6.1.2 Redundant Arrays of Independent Tape (RAIT) drives

[19] RAID: Theory and Practice, Harold Henry Chaput, May 8th, 1997. http://www.cs.utexas.edu/users/chapu

RAIT is similar in concept to RAID, except that the medium is magnetic tape instead of magnetic disc. The advantages of RAIT are higher data throughput, error correction, and the increased capacity of more than one tape. The disadvantages of RAIT are that groups of tapes must be kept together since the data is spread across the tapes. If compression is used, tapes within the group may fill at different rates, causing problems in using total tape capacity.

In general, RAIT is useful where very high data transfer rates and recovery from loss of media is desirable. Since RAIT storage is a linear device, random access time will be much slower than RAID.

6.1.3 Storage Area Network (SAN)

As more and more data are maintained electronically at EPA and as EPA partners deliver more information electronically as part of the REI initiative, then the volume of electronic information may challenge existing storage architectures and the paths for moving this information. One possible solution to the need for high-speed transfer between multiple storage devices is a SAN.

A SAN is a managed high-speed network that provides any-to-any interconnection of server and storage elements.[20] SANs are based on a "fabric"of Fibre Channel (see Section 6.1.4 below) hubs, switches, and gateways connecting storage devices such as disk arrays, optical disks or tape libraries to local or wide area network servers. By separating information storage and management from information processing, a SAN provides the flexibility for storing data on many different devices and making information readily available across the enterprise. A SAN can be viewed as an "information utility" that provides a faster, more effective way to deal with rapidly increasing volumes of information while improving backup and recovery capabilities and enabling higher performance and storage availability. "Storage Area Networks offer a flexible alternative to server-specific storage and large enterprise arrays, which are inherently expensive and limited in scalability."[21]

The following are some of the specific advantages of a SAN:

- Capacity and performance are both scalable, allowing incremental expansion as the need for greater storage and faster access arise.
- Companies can make storage decisions independent of server considerations.
- Storage management is removed from the server, freeing CPU resources for running business applications.
- Data duplication can be substantially eliminated.
- Data availability is no longer tied to the availability of a single server.

---

[20] Storage Area Networks: Putting Data to Work for e-businesses, IBM Technology Group, Copyright IBM 1999.

[21] SAN, Storage Area Networks; Copyright Data General Corporation 1999

- Centralized storage management can simplify administration, reduce long-term operating costs and allow storage personnel to focus on strategy and planning rather than day-to-day problem solving.

6.1.4 Fibre Channel[22]

Fibre Channel is a new technology for communicating with storage devices that can be used to improve speed between servers and high-speed storage. Fibre Channel technology allows higher input/output (I/O) performance and higher data transfer speed (~106 megabytes/second) than other data transfer technologies such as UltraSCSI (~40 megabytes/second). Fibre Channel is significant because it allows storage devices to be geographically isolated from the network servers, or to be part of a SAN.

With Fibre Channel connectivity, communication between hosts and devices does not have to be done directly. Hubs and switches can be used to create Fibre Channel storage area networks, which can connect multiple hosts with storage devices. Fibre Channel supports fiber optic cabling, which is not affected by electronic noise and can be run 500 meters between devices and up to 10 kilometers in the future. Because of this, users can store their information far away from its source, making the storage solution immune to computer center calamities.

Up to 126 devices can exist on a Fibre Channel connection. This means that adding capacity with additional storage elements is relatively easy. Small Computer System Interface (SCSI) buses, on the other hand, are limited to 16 connections.

Finally, Fibre Channel is designed to allow non-disruptive removal and addition of storage devices, so that devices can be swapped without affecting the flow of information on the data path.

6.1.5 Jukeboxes

Jukeboxes are named after the Wurlitzer jukeboxes used to select and play 45 rpm records, and they function in a similar fashion. In a storage jukebox, electromagnetic robotics are used to manipulate the storage medium, placing them in predetermined racks or pigeon holes, and inserting or ejecting them from the storage drive or drives.

In the old Wurlitzer jukeboxes, there was only one drive and so only one disc was playable at a time. In modern storage jukeboxes, adding more drives allows simultaneous access to more than one disc or tape. However, having more than one drive introduces a complex problem in queuing theory so a management system is needed to optimize access.

---

[22] Fibre Channel information for the CLARiiON community; Copyright Data General Corporation 1999

Jukeboxes come in many different sizes. Small ones may have one drive and ten slots for media. On the high end, optical or tape juke boxes will store hundreds of discs or tapes providing terabytes of near-line data (see Table 6-1 below).

6.1.6 Storage Systems

Most major disk and tape vendors have jukebox products on the market. The following table compares low-end and high-end storage systems using magnetic cartridge tape, CD-R/CD-ROM, DVD-RAM, and Magneto-Optical Storage.

The table uses recent price quotes for the storage system, the storage system management software, and the cost of the media to compute a dollar cost per megabyte for a fully populated unit. Transfer speed is also shown to characterize throughput, a consideration that is important for media refresh and backup.

The storage systems in Table 6-1 are listed in increasing order of cost per Megabyte.

Table 6-1. Optical and Magnetic Storage Devices[23]

Make/ Model/ Media	Interface/ Transfer Speed	# Drives/ # Media/ Total Capacity	Cost of Unit/ Cost of Media*	Overall Cost/ Cost$/Mbyte
Exabyte 280 Library 8mm Tape	SCSI-2 Fast 12 MBytes/sec**	80 2 1.6 TBytes **	$28,000 *** $70/tape	$33,600 $0.0200
Plasmon DV240-04DR DVD-RAM	SCSI-2 1.38 MBytes/sec	240 4 638.4 GBytes	$24,000 *** $35/disk	$32,400 $0.0496
Plasmon MOD-520 MOD	SCSI-2 2048 KBytes/sec	104 4 540 GBytes	$30,000 *** $89/disk	$39,256 $0.0710
Plasmon D120-22 Library CD-ROM/CD-R	SCSI 2 3600 KBytes/sec ROM 600 KBytes/sec R	120 4 78,720 MBytes	$14,395 *** $6.75/disk	$15,205 $0.1932
Panasonic (Single) LF-D101U DVD-RAM	SCSI-2 Fast 1.39 MBytes/sec	1 1 2.66 GBytes	$679 $35/disk	$714 $0.2621
Arch Data CD NetStor C100 CD-RW	Fast SCSI 2 1200 KBytes/sec R 600 KBytes/sec W	6 (4 RO, 2 RW) 100 65,600 MBytes	$21,700 *** $10/disk	$22,700 $0.3460

[23] Pricing from Arch Data Systems, Inc., www.cdstorage.com, 800-640-0530.

Make/ Model/ Media	Interface/ Transfer Speed	# Drives/ # Media/ Total Capacity	Cost of Unit/ Cost of Media*	Overall Cost/ Cost$/Mbyte
Panasonic (Single) CDR7502B CD-R	SCSI 1200 KBytes/sec R 600 KBytes/sec W	1 1 656 MBytes	$249 $6.75/disk	$256 $0.3902

* Quantities of 100
**Specifications are for uncompressed data; use of compression gives better values and yields lower cost.
*** Price includes jukebox management software which is typically 50% of overall price.

6.1.6.1 Hard Disk (HD) Caching of Storage Systems

Hard disk caching is the process of mirroring (copying) CD-ROM or DVD discs to a hard disk co-located in a jukebox. The data that have been copied are now accessible directly over the network from the hard disk, rather than from the CD itself. The primary advantage of this is that the HD cache can keep a directory of all the entries (files) on the jukebox, allowing for the ability to browse all discs in a changer without having to cycle through them to read the directories. This directory provides for easy verification of the access request and the location of the correct disc in the jukebox. The secondary advantage of this is access speed, particularly when the CD drive has stopped spinning or has "idled down" because it has not been used for a period of time, or because a different disc has been loaded onto the drive. [24]

Caching CD and DVD data to the local hard disk can yield significant performance improvements in the following cases:

- Reaccessing recently accessed data
- Access to data on multi disc CD changers
- Access to data from more CDs than the tower has drives
- Accessing multiple files from multiple discs

[24] http://www.cdstorage.com/towers/NetStorC100.htm, Copyright c 1993-1999 Axis Communications, Co 1998-99 Arch Data Systems, Inc. All rights reserved. Revised: June 14, 1999.

6.2 Storage Media

Electronic record storage is possible today on a variety of storage media. Twenty years ago, the predominant storage medium for large quantities of electronic data was 9-track magnetic tape reels. Today there are other choices such as semiconductor memory, removable compact disc, and tape cartridges.

Semiconductor memory includes, but is not limited to, Read-Only Memory (ROM), Programmable ROM (PROM), Random Access Memory (RAM), Charge Coupled Devices (CCD), and Bubble Memory. These devices have very fast access times but relatively low capacities for data storage. They are also very volatile; some require constant electricity to maintain the stored data. Because of their low storage capacity and low permanence, this type of media is not currently a viable choice for large data or long term storage.

Rotational devices include a variety of disc devices such as magnetic disk, optical disks, and magnetic storage systems. They are all based on creating a long-lasting semi-permanent "engraving" of digital bits on a circular substrate (disc). Examples of these devices are magnetic hard drives, CD-ROM discs, removable disk cartridges, and DVD.

Linear devices for storage are primarily magnetic tape that stores digital bits on a magnetic ribbon that could be hundreds or thousands of feet long. While the storage is long-lasting, it is not as easily refreshed as data on rotational devices. Examples of this medium are 9-track tape reels, Digital Audio Tape (DAT), and Digital Linear Tape (DLT).

This study does not examine memory devices due to their relatively small storage capacity. The following sections further explain the other choices.

6.2.1 Disc Storage Media

Media for rotational storage devices are listed below:

- **Hard Drives (HD) Magnetic Disk.** Single and multi-platter rigid magnetic disks operating in a vacuum-sealed environment impervious to dust and humidity.

- **Floppy Drive (FD) Magnetic Disk.** Single platter flexible magnetic disks in permeable cartridges exposed to dust and humidity.

- **CD-ROM.** Compact plastic disc that can only be read from. The digital information on this type of disc is usually injection molded into the substrate against an aluminum reflective coating.

- **CD-ROM XA.** Compact plastic disc. The "XA" stands for Extended Architecture. CD-ROM XA is generally consistent with the ISO 9660 logical format but designed to add better audio and

video capabilities so that a CD-ROM can more easily be used for multimedia applications and Photo CD discs.

- **CD-RAM.** Compact plastic disc that can be read and written to.

- **CD-R.** Compact disc-recordable.

- **CD-RW.** Compact disc-rewritable. The most recent addition to the compact disc family. It was originally called “CD-Erasable.” The official name is CD-ReWritable, and it is a media and recording system that allows the user to erase previously recorded information and then record new information onto the same physical location on the disk.

- **CD-WO.** Compact Disc Write Once. A CD-ROM version of the WORM (Write Once Read Many) technology. Format for mastering and replication. CD-WO discs conform to ISO 9660 standards and can be played in CD-ROM drives.

- **DVD-RAM.** A new type of Rewritable compact disc that provides much greater data storage than today's CD-RW systems. The caddy-mounted discs will initially provide 2.6GB per side on single or double-sided discs.

- **DVD-ROM.** The read-only format supports discs with capacities of from 4.7GB (enough for an MPEG-2 compressed full-length movie) to 17GB and access rates of 600KBps to 1.3MB/s. Backward-compatible with CD-ROMs.

- **DVD+RW.** The name used for the rewritable DVD standard being promoted by Hewlett-Packard, Philips and Sony (to rival DVD-RAM) before the DVD-Forum insisted on it being changed to '+RW'.

- **MOD.** Magneto-Optical Disk. A plastic or glass disk coated with a compound (often TbFeCo) that has special properties. The disk is read by shining a low-intensity laser onto the media and examining the polarization of the reflected light. The disk is written by shining a higher-intensity laser to heat the material up to its Curie point, where it becomes susceptible to a magnetic field. When the media cools again, its state is “frozen.” The polarity of the reflected light during a read depends on the polarity of the magnetic field under which the media was last cooled.[25]

[25] http://alumni.caltech.edu/~rdv/comp-arch-storage/FAQ-1.html

6.2.2 Tape Storage Media

Tape media is probably the most widely used long term digital storage due to the lack of other digital storage options between 1960-1985. In general, tape media have worked well for the computer storage needs of government agencies and private corporations; however, problems are beginning to surface with reading older tapes as well as the lack of availability of large tape drives.

Most new tape storage technology is centered on much smaller size, higher density tape cartridges. On the high-end are the 3490 cartridges used in robotic tape silos. On the low-end are a variety of small tape cartridges used primarily as backup and restore systems, but are also now part of sophisticated array systems.

Tape media are adversely affected by heat, cold, humidity, handling (tape can be creased or stretched) and magnetic fields generated by motors, magnets, or scanning devices. Tape media are also susceptible after long periods of storage to "print-through," which is the phenomenon of magnetic imprints on one layer of tape transferring to another layer of tape.

Media for linear storage devices are listed below:

- **9-Track Magnetic Tape**. Half-inch tape on open reels that come in various sizes and densities. Open reel tapes from the 1970s will have lower densities. Newer tapes have densities of 6250 bits per inch (bpi) or greater. Open reel tapes are more subject to dust and humidity than cartridge tapes.

- **3490 Magnetic Tape Cartridge**s. Wide tape format with 36 read/write tracks can store up to 800 MB of uncompressed and 2.4 GB of compressed data per cartridge. New high-end linear tape cartridges can store 10-20 GB uncompressed data and 30-80 GB of compressed data.

- **Digital Audio Tape (DAT)**. High-quality 4mm magnetic tape in a cassette, with capacities up to over 1 Gigabyte used mainly as an archival and back-up medium. It is sometimes used as a transfer medium for CD-ROM audio.

- **Digital Linear Tape (DLT)**. Tape cartridge with a high storage capacity (10-20 Gbytes). It is used in backup and restore systems and in tape array systems.

- **8mm Tape**. Cartridge with 8mm-wide tape and a storage capacity of 5 gigabytes (GB) or more, usually used for digital storage and in 8mm video cameras.

- **4mm Tape**. Cartridge with 4mm-wide tape and a storage capacity of 2 gigabytes (GB) or more.

6.3 Shelf -Life of Storage Media

There are considerable differences of opinion among experts on the life expectancy of storage media. A fairly extensive literature search has provided differing data. Table 6-2 provides information from a National Media Lab Report published in March 1995.

The shelf life of media is affected by environmental factors such as dust, temperature, humidity, and handling. Magnetic media such as tape are susceptible to magnetic fields as well. High temperatures will weaken the magnetic imprint (orientation of magnetic particles) on the tape. Humidity can cause the layers in the tape to separate, and poor handling can crease or stretch the tape affecting the readability. In addition, magnetic tape can experience a phenomenon known as “print through” which occurs when adjacent layers of tape on a reel transfer some of their magnetic imprint to the other layer of the tape if the tape is stored and left unused for a long period of time. Fortunately, some tape systems store error correction data on the tape so that marginally damaged data are recoverable.

Optical discs appear at present to be much more durable than magnetic tape and most are not affected by magnetic fields. High humidity is the major cause of damage to optical disks, as this can cause corrosion of the reflective layer of the disc. Exposure to very high temperatures (greater than 65 degrees Centigrade) can cause the plastic substrate to warp, which would destroy the ability to read the disc. One advantage of optical media is that 3 tracks are used for error correction data which allows data recovery even when the discs generate data errors. (See section 6.3.2).

Table 6-2. Life Expectancy of Storage Media[26]

Data Storage Product	Format/Technology	Life Expectancy (Years)
Magnetic Tape	3480/3490 cartridge Digital Linear Tape (DLT) 19mm digital data format (DD-2) Quarter-Inch Cartridge (QIC) D8 (Data 8mm)	10-30 10-30 10-15+ 5-30 2-30
CD-ROM	Yellow Book	5-100+
Magneto-Optical	3.5" or 5.25"	5-100+
WORM (Write Once Read Many)	Pits on bimetallic film Ablative pits on tellurium Thermal bubble Phase change metallic alloy Pits on organic dye/polymer	100 30-40 30 10 10
Microfiche	Photographic Image on Film	12-100+
Paper	Ink Printed on Paper	25-500

6.3.1 Life Expectancy of Magnetic Tape

"Experience indicates that physical lifetimes for digital magnetic tape are at least 10 to 20 years .... One government agency responsible for maintaining meteorological data archives recently transcribed approximately 20,000 ten-year-old 3480 tape cartridges, of which only two cartridges had unrecoverable errors. Properly cared for reel-to-reel, 9-track computer tapes recorded in the 1970's can still be played back in the 90's, even though the 9-track format became obsolescent in the 80's...."[27]

Network server storage may use Data 8mm (D8) tape medium for backup and archival purposes. The life expectancy of this medium may be as little as 3 years according to some practical experiences (see http://www.cclabs.missouri.edu/~ccgreg/tapes.html). The reasons for a shorter than expected life are humidity and the number of passes the tape makes over the read/write heads of the tape drive. IBM recommends that a tape be used for no more than 500 passes before it becomes unreliable.[28]

In summary, 9-track reel tape has longer life expectancy and is more durable than the smaller tape media, particularly the 8mm tape cartridges. The upper limit of life expectancy for tape is around 30 years, but the working limit on tape that has many passes is much less. Magnetic tape is a viable

[26] National Media Lab Report published in March 1995, 3M Corp.

[27] Dr. John W. C. Van Bogart, National Media Laboratory, 3M, St. Paul, Minn.

[28] Chapter 9 in the V2R2 AS/400 System Operator's Guide, SC41-8082

medium for archival if the retention period is less than 30 years, the archive process does not excessively use the tape, and if the storage conditions are low in dust and humidity.

6.3.2 Life Expectancy of Optical Discs

Recently considerable effort has been expended on trying to determine the life expectancy of optical media. Test labs use the following two measures:

- **BLER** (BLock Error Rate). This is the "raw" digital error rate before any error correction (intrinsic to the disc) is applied.

- **BLERmax**. The maximum number of BLERs allowed on a disc. According to the industry standard, a CD-ROM is allowed a BLER of up to 220 before it is considered a "bad" disc.

Studies of how BLERmax varies with temperature and humidity have concluded with 95% confidence that 95% of a test population of rewritable CD Media (CD-R) will have a data lifetime of greater than 217 years if stored in the dark at 25°C, 40% Relative Humidity after being recorded. The studies used a "middle range" of 50 for BLERmax. If the industry standard specification of BLERmax less than or equal to 220 had been used as the end-of-life criterion, then this same analysis would predict that with 95% confidence, 95% of the test population of CD-R discs will have a data life of greater than 12,000 years.[29]

The above study also showed that the stability of CD-R (rewritable) media was good even at temperatures of 60EC (140EF) and that normal CD-ROM media did not display good stability over time at this temperature. This is partly because CD-R media use a gold reflective layer while CD-ROM media use an aluminum reflective layer and because of the method in which data is imprinted.

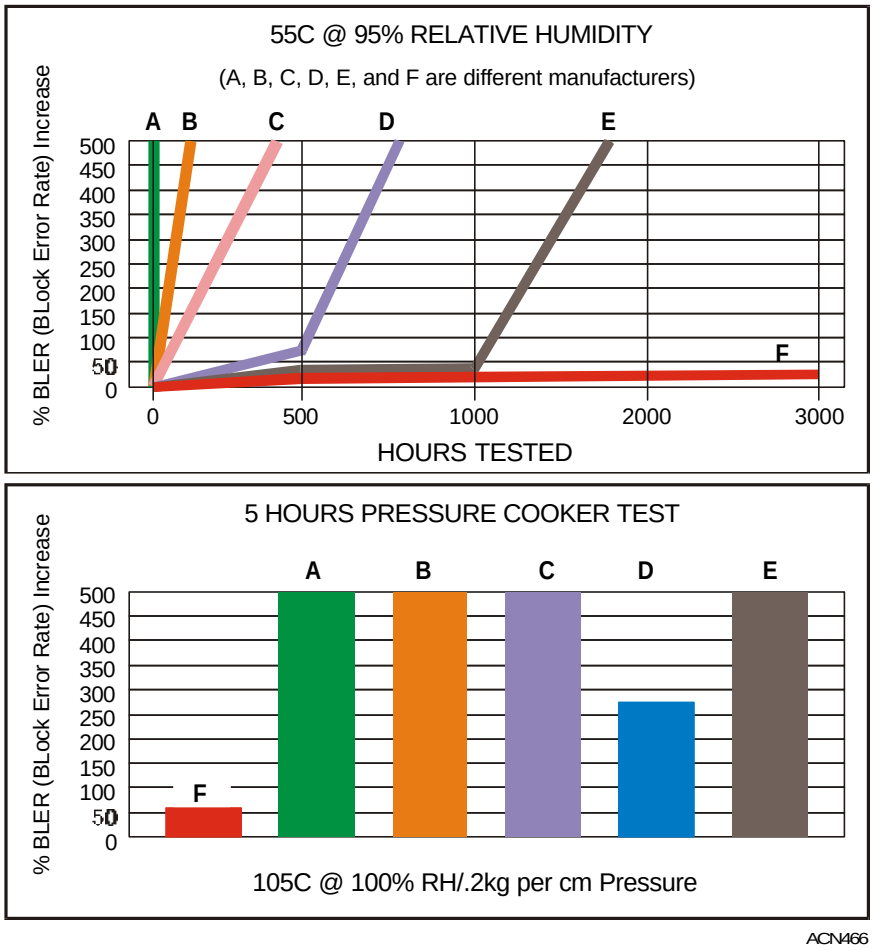
The reliability and shelf life of CD media at less than optimum environmental conditions will vary greatly depending on the manufacturer of the CD media. Briefly, not all CD media are created equal. This is because the functionality of CD media depends on the quality of the polycarbonate substrate (plastic base), the reflective coating, the recordable dye, and the stamping or recording process.

For instance, the reflective coating on standard CD-ROMs is made of aluminum, which is corrosive. CD-Recordable media use either a 24 kt gold or a silver coating. The gold coating is non-corrosive and so is much more stable.

Figure 6-1 illustrates the failure rate of optical discs from different manufacturers.

[29] http://www.cd-info.com/CDIC/Technology/CD-R/Media/Kodak.html

Figure 6-1. Life Expectancy of Optical Disks by Manufacturer



Comparison of the life expectancy of six vendors' products shows that high-quality optical discs should be used for data that will be archived in excess of five years. The quality of the discs are determined by the manufacturing quality control as well as the type of reflective coating used. A 24kt gold coating is one of the better choices.

In summary, good quality optical discs will last in excess of 200 years at room temperature and low humidity which makes them ideal for the retention of digital data.

6.4 Pros And Cons of Storage Media

The alternatives for storage media accepted by NARA are paper and open reel tape. This may soon change to include cartridge tape and optical disk when NARA releases their guidance for electronic records submitals. EPA can, however, choose whichever storage medium meets the cost/benefit and operational needs of the Agency for internal records archival.

Table 6-3 summarizes the pros and cons of the various storage media alternatives:

Table 6-3. Comparison of Storage Media

Media	Mechanism	Pros	Cons	Portability	Longevity	Capacity	Performance	Cost/MB*
**Magnetic Disk (disk array)	Magnetized Particles	- proven reliability - cost effective - fast access time	- susceptible to magnetic interference (i.e. power surges, metal detector, electric fields) - slowly degrades over time - not very scalable	Poor	5-10 years	Variable	<=20MB/s	$0.04 (avg.)
Optical Disc	Substrate Ablation	- can store large amounts of data on a single disk - scalable - immune to magnetic fields	- susceptible to scratches on the media surface - many differing formats risk future compatibility	Very Good	CD: 5-30+ yrs. DVD: 40+ yrs.	CD: 650MB DVD:4.7- 17GB	CD: 4.8MB/s (32X) DVD: 8.1MB/s (6x)	$0.05 - $0.40
Magnetic Tape	Magnetized Particles	- proven reliability - scalable - cost effective - fast data transfer speeds	- susceptible to magnetic interference (i.e. power surges, metal detector, electric fields) - slowly degrades over time - media easily wrinkled or damaged	Very Good	DLT :10-30+ yrs. QIC: 5-30 yrs. 8mm: 2-30 yrs. 3480: 10-30 yrs.	- 10-20GB+ U‡ 20-40GB+ C‡ - 80MB-2GB U 160MB-4GB C - 5GB+ U 10GB+ C - 0.8GB+ U 2.4GB+ C	- 1.5-20MBps - 62.5KBps-70MBpm - 3MBpm-120MBpm	- $0.003- $0.
Microfilm/ fiche	Photo-Chemical	- proven reliability - scalable - longevity	- must be created by outside service - special equipment needed to read - poor interface with computer equipment	Very Good	12-100+ yrs.	Variable	N/A	N/A

Media	Mechanism	Pros	Cons	Portability	Longevity	Capacity	Performance	Cost/MB*
Paper	Ink printing	- proven reliability - scalable - longevity	- bulky - difficult to find/access specific documents - not easily refreshable	Good	25-500 yrs.	Approx. 2.4K per page	N/A	N/A

* Based on a 1997 Gartner Group study ** Floppy Diskettes were not considered due to their low storage density ‡ U= Uncompressed ‡ C= Compressed

6.5 Technologies for Image Processing

The Department of Defense (DoD) and NARA are studying the conversion of paper records to electronic images of these records for long term storage as an alternative to storing the paper. In 1996, the DoD initiated a study of digital image format standards. DoD sought and received NARA's collaboration in this effort, hoping to ensure that NARA's preservation needs and DoD's operational requirements are in tandem. The results of this continued study were just released by DoD on June 1, 1999 as Report Number GA22F042, *Electronic Digital Imaging Standards for Archiving Records*[30]. This study is a definitive analysis of the topic and is worth consulting for an in depth understanding.

Findings pertaining to electronic imaging from the above report and from the NARA website (http://www.nara.gov) are summarized as follows:

- The DoD study team focused on the imaging of personnel records, manuals, standards, directives, and documents scheduled for declassification. In their survey of 35 government agencies of which 25 responded, 92% of respondents believed that NARA or OSD (Office of the Secretary of Defense) should provide standards on how records should be stored in an electronic archive, and 88% requested direction on how to establish, implement, and maintain an electronic archive.

- *There was widespread recognition in the study that if electronic records of documents are to be maintained, then the optimal solution is to store the electronic files as an electronic image.* This image should be an exact replication of the original document and is the most appropriate method for the long term preservation of a record. NARA is concerned that electronic records stored in application-specific formats such as Microsoft Word and WordPerfect may not be readable (due to version problems) in the future.

[30] *Electronic Digital Imaging Standards for Archiving Records*, Final Report, June 1, 1999 written for the Department of Defense under Contract GS35F4863G by Susanne H. MacTavish and Michael R. Pickard of Lockheed Martin Technical Services, Falls Church, Virginia.

- Scanning is the most common method of converting hardcopy documentation into an electronic archive.

- There are 33 image formats (e.g., BMP, BIIF, GIF, IMG, JPEG, PCX, PDF, PICT, TIFF, WPG, etc.) in use. The most common imaging format used to archive imaged documents is Tagged Image File Format or TIFF. Portable Document Format (PDF) is also an accepted standard. Joint Picture Experts Group (JPEG) is the most widely used compression standard for photographs.

- The DoD study team survey found that the use of optical disks by government is very high. This is of great concern to NARA because most optical disk formats (except CD-ROM) are not standardized. NARA fears receiving optical disks for which hardware to read will not be available.

- Organizations should plan and budget for migrating digital images to the most current storage technology every 3 to 5 years.

The general consensus at present is that converting to electronic image format (e.g., TIFF, PDF) represents the best solution for access to archived records in the future, primarily because application-specific formats (such as WP5, WPD, WP8, DOC, PPT, XLS, WK4) may not be readable in the future without the proper application software, system software, and hardware. Converting all records to a single (or, at the most, two) image format also makes the task of migrating to the most current technology a more manageable and relatively simple task.

The conclusion of the DoD study is that the latest TIFF format should be used for archiving records, and that the PDF format should be used to disseminate archived records.

TIFF is the most common format for storing digital images since almost every scanner on the market today is capable of creating a TIFF file that is an exact replica of the scanned document. There are many TIFF versions; the latest specification is version 7.0.

PDF has become the de facto standard for publishing documents on the World Wide Web (WWW). The main advantage of PDF is that it is a device-independent format. PDF files typically occupy less space than the original document format, and PDF files

produce an exact replica of a document including figures, pictures, and tables. Although PDF is a proprietary format of Adobe Corporation, they have made it available to other vendors.

6.5.1 Costs of Imaging

The DoD report cites the following costs for electronic imaging:

Method of Conversion	Format	Cost
Single pages to electronic image	TIFF	10-16 cents/page
Bound book pages to electronic format	TIFF	22-25 cents/page
File image to electronic image	TIFF	10 cents/page
Fiche image to electronic image	TIFF	12 cents/page
Single page or TIFF image to PDF	PDF	20 cents/page

These costs assume a large volume of conversions and that equipment costs, acquisition and maintenance, staff training, quality control and software costs are spread over a large business base.

6.5.2 Census 2000 As a Case Study

"The new millennium's first major census will take place in [the year] 2000, and for the first time in its history, the Census Bureau will rely on technology to read entire survey response forms. The bureau's ambitious goal is to process 1.4 billion forms in 100 days, with an accuracy rate of 98 percent.

High-speed scanners will convert the forms into electronic images. The images then will pass through a gauntlet of software applications that will extract data by reading bar codes, check marks, handwriting and printed text. Finally the data will be automatically filed as database records.

Proponents of imaging technology (hardware and software that convert paper documents into computer-compatible digital versions) see a lot riding on the $49

million Data Capture System 2000 project, which will tackle one of the country's largest paper processing tasks." [31]

The results of the Census Bureau system will present a good case study in document imaging technology which should be referenced at a future date.

[31] *Imaging Tools Cut the Paper Trail*, by Luba Vangelova, Government Executive Magazine http://www.govexec.com/tech/articles/1297tech1.htm

7 STATUS OF STORAGE SOLUTIONS AT EPA

7.1 Mainframe Storage

EPA currently utilizes Automated Tape Libraries (ATLs) or silos from StorageTek for long term storage and archival of digital information. The silos use square tape cartridges that are stored and retrieved with robotics. IBM's Data Facility System Managed Storage (DFMS) product provides the software to operate and manage the tape archival process. This software currently controls 1,250,000 data sets (12 terabytes) for over 1,344 active accounts.

There are several different types of data sets, but in general, data sets that have not been accessed in 14 days are migrated to tape and retained for 1 year.[32] After 1 year of near-line storage, the user is notified that the data will be destroyed and the user may request that the data sets be archived for 2 to 7 years according to EPA Operations Policy No. 210.09. These long term archival tapes are stored in on-site and off-site tape libraries. The off-site library is dedicated to critical overhead and disaster recovery type data. The on-site libraries house both overhead and customer data.

These systems are well thought out and work very well, and have established policies and procedures to manage the archival process.

7.2 LAN Storage

EPA is currently using ARCserve software to interface with a Small Computer System Interface (SCSI) Adic FastStor Digital Linear Tape (DLT) subsystem for LAN based data backup and archival. The storage subsystem holds 247 GB of data (roughly seven days of backups) on seven DLT cartridges. Once all seven cartridges are filled they are removed, and stored in the off-site storage library. While the tapes are still on-line, users have the ability to perform their own file restores. Once the tapes have been moved to the storage library, a 24-hour retrieval time is imposed.

[32] *EPA-NCC IBM Storage Management Plan* dated August 25, 1999

Differential backups (files that have been changed since the last backup) are stored for 60 days. Weekly full backups are stored for 6 months, and the monthly full backups are stored for a period of one year.[33]

7.3 Distributed Systems Storage

The current Unix based network employed by EPA calls for each server to be configured with local disk storage except for central UNIX servers. These central UNIX servers use storage from a Clarion RAID system and are backed up using StorageTek's Central Archive Manager (CAM). CAM utilizes EPA's IBM mainframe as the backup server with one of StorageTek's automated tape libraries for storage. Once data are transferred to tape cartridges, the procedures for archival described in the Mainframe Storage section above are followed.

7.4 Image Processing

The most widely used image processing system within the Agency is the Superfund Cost Recovery Image Processing System (SCRIPS). SCRIPS is an application used to prepare cost recovery packages used for litigation and cost recovery. Through the use of image processing technology, SCRIPS facilitates the storage and retrieval of all site-specific Superfund cost documentation. Input records include timesheets, travel vouchers, miscellaneous expenditures, contract invoices, IAG payments, and other supporting cost documentation.

SCRIPS captures financial documentation images at Regional and field sites (i.e., Washington, Cincinnati, RTP) throughout the country. Initially, all those images were stored centrally at EPA's National Computer Center (NCC) data processing facility. The system has since been decentralized into each of the Regions. SCRIPS is under the control of the Superfund Accounting Branch (SAB) of the Financial Management Division (FMD) in the Office of the Comptroller (OC) under the Office of Administration and Resource Management (OARM).

SCRIPS is just one example of image processing technology in use at EPA. In planning for the implementation of an Agency-wide electronic records management capability,

[33] *EPA-NCC LANS Storage Management Plan* dated August 24, 1999

the integration of pre-existing image processing capabilities must be considered. The selected records management architecture should minimize the degree of change while maximizing the reuse and preservation of past investments.

8 STORAGE AND INFORMATION MANAGEMENT

Traditionally, the term "records management" has been used to describe the management of the media on which information is recorded. Managers are now using the term "information management" to describe the processes necessary for the creation, use, and disposal of information regardless of the media on which it is recorded.[34] Other terms in prominent use in government are "electronic records management" or ERM and "Record Management Applications" or RMA. This section discusses the technology of storage management and information management.

8.1 Storage Systems Management Technology

The management of storage systems develops the idea of making storage a true subsystem of a computing environment that can be accessed not just from one computer but from many. This requires separation of hardware so that the storage unit(s) have their own power systems and bus architectures, and requires the logical separation of the storage so that a particular storage device is not a "device" to the computer but a "volume" of information that has a name related to the information and not to the device. The latter is sometimes referred to as "virtual file storage." Virtualization frees the computer and the storage management application from decisions involving hardware and allows the application to manage logical constructs such as Volume Names and Directories. For instance, in a virtual storage system, a Directory may be spread across more than one storage device.

In addition to the SAN and Fibre Channel technologies described in Section 6.1, management of storage systems includes the following additional technologies.

8.1.1 Hierarchical Storage Management (HSM) [35]

Hierarchical Storage Management (HSM) is a strategy that places data on different types of storage in accordance with the frequency of access to the data. Data that are

[34] http://www.epa.gov/irmpoli8/recmgmt/chaptr1.txt.html

[35] http://alumni.caltech.edu/~rdv/comp-arch-storage/FAQ-2.3.html

accessed frequently are kept on high-availability[36], fast access, limited capacity devices. Data that are accessed infrequently are kept on medium-availability, medium access, high capacity devices. Data that are archived are kept on low availability and slow access storage. Key to the concept of HSM is to keep data with less demand on cheaper media to lower the overall costs of storage. "An effective HSM implementation can increase the total amount of storage that can be managed in a single installation, as well as presenting to the user the illusion of infinite disk space."[37]

HSM systems transparently migrate files from magnetic disk to optical disk and/or magnetic tape, usually robotically accessible, following a pre-determined period of inactivity. Then when a file is required by a user, HSM transparently moves it back to hard disk for fast access.

HSM systems may offer integrated backup to keep more than one copy of a file for data reliability. They may also offer integrated migration from other systems such as file servers to the central storage location, and then to a robotics library.

The StorageTek silos attached to EPA's mainframe system and the associated management software, including CAM, is an example of an HSM system.

8.1.2 High Performance Storage Systems (HPSS) [38]

The High Performance Storage System (HPSS) is new, open system software designed to manage petabytes of data produced and used by supercomputers. The primary objective of HPSS is to move very large data objects between high performance computers, workstation clusters, and storage libraries at speeds many times faster than is possible with today's software systems. For example, HPSS can manage parallel data transfers from multiple network-connected disk arrays at rates greater than 1 Gbyte per second, making it possible to deliver high definition digitized video in real time.

---

[36] Availability refers to how accessible is a storage medium . For instance, on-line devices have high availability, even though access speed will vary from slow to fast, depending on the device.

[37] http://www.dlttape.com/resources/platformguide/dlt_sol_terms.htm

[38] http://www.sdsc.edu/hpss/hpss1.html

HPSS is designed to make data immediately available to all networked computing nodes regardless of size or distribution. HPSS is an open system based on the IEEE (Institute of Electrical and Electronic Engineers) Mass Storage Reference Model Version 5, the established design guide for very large scale storage systems. All computer and storage nodes may be attached directly to the network so that data are transferred by the most direct route, at network speeds, without interruption.

HPSS is a major development project which began in 1993 as a Cooperative Research and Development Agreement (CRADA) between government and industry, and is now nearing its third major release. The HPSS collaboration is based on the premise that no single organization has the experience and resources to meet all the challenges represented by the growing imbalance in storage system I/O, capacity, and functionality imbalance. Over 20 organizations including industry, Department of Energy (DOE), and other federal laboratories, universities, and National Science Foundation (NSF) supercomputer centers have contributed to various aspects of this effort.

HPSS is designed to deal with the massive amounts of data (from multigigabyte-to-terabyte datasets found primarily in a supercomputing environment), to assist with the user need for a high-level application-centered view of information, and to integrate distributed information repositories.

8.2 Management of Information

There are many records management applications available today to manage the workflow and disposition of company information. Some packages integrate better than others with the systems in place in an organization. Some packages have greater capability in workflow control, while others are better suited for managing storage.

Selection of a records management application will involve a careful assessment of business processes, system integration, information flow, information volume, and management objectives. Such an assessment is beyond the scope of this paper, but must be conducted before selecting a product and going through the expense of integrating it into the organization's IT infrastructure.

On April 11, 1997, DoD reissued directive 5015.2 (see Appendix C) which defines a list of requirements for records management. The directive has been accepted by

NARA as guidance in this area. The directive requires the Defense Information Systems Agency (DISA) to establish and maintain the capability to test and evaluate records management applications that are seeking DoD approval and lists criteria in broad terms. As a result of DoD's involvement and NARA's acceptance, this directive has become an ad hoc standard for Records Management Application (RMA) software (see Section 9).

8.3 Migration of Information

A recent report done for the Department of Defense stated the following with regard to migration of storage archives:

"As the operating environments of digital archives change, it becomes necessary to migrate their contents. There are a variety of migration strategies for transferring digital information from systems as they become obsolete to current hardware and software systems so that the information remains accessible and usable. No single strategy applies to all formats of digital information and none of the current preservation methods is entirely satisfactory. The costs of these strategies also vary greatly. The general rule of thumb places migration costs at between 50-100% of the cost to create the original digital image of the document."[39]

What this means for EPA is that over a period of time, regardless of the hardware and software combination used for storing digital archives, they will eventually become obsolete and the existing data must be transferred to current hardware and software platforms. This makes the choice of a digital storage format of paramount consideration. The format must be an existing standard, and must be projected to remain a standard for some time to come. That format must be easily transportable between hardware platforms. For text and image documents that will probably mean utilizing TIFF or PDF formats. It is critical for future migration that uniform formats be selected for each type of information being archived.

[39] *Electronic Digital Imaging Standards for Archiving Records*, Final Report, June 1, 1999 written for the Department of Defense under Contract GS35F4863G by Susanne H. MacTavish and Michael R. Pickard of Lockheed Martin Technical Services, Falls Church, Virginia.

Use of one of the evolving interchange standards, such as the Basic Image Interchange Format (BIIF) or Electronic Document Interchange Standard (EDIS) allows for the conversion of images in many different types of formats to a single robust format. EPA would do well to commission a further study into the use of these interchange standards.

8.4 Cataloging, Metadata, and Indexing

To provide for viable access to records, the RMA has to provide for information cataloging, metadata, and indexing. Cataloging of records will include information on the type, quantity, location, and retention period. In addition to this, cataloging may include metadata and/or indexing.

Metadata is more detailed information on the record and may include fields for subject, author, organization, data sources, key words, creation date, retention period, usage constraints, language, cross-references, etc. The standard for bibliographic data is US MARC and the AACR2 (Anglo American Cataloging Rules, Second Edition). There are additional cataloging initiatives for archival of electronic records such as the Encoded Archival Descriptor (EAD), the Text Encoding Initiative (TEI), and the Resource Definition Framework (RDF) for XML. There are metadata standards for the header areas of TIFF and BIIF files, and for many other file formats.

Indexing of records involves recording a link to the record based on information that is part of the record. Indexes can link to every word in a record or to just the keywords, or to the metadata for the record. The main purpose of indexing is to rapidly locate a specific record or to generate a list of records that meet certain criteria. For example, indexing can facilitate retrieval of all records stored over several years relating to a particular subject.

With cataloging, records that have exceeded their retention period can be easily located for removal. Metadata can help managers decide if the retention periods of certain records should be extended. Prior to destruction of records, managers may be notified and given a period of time in which to prevent the destruction by extending the retention period.

8.5 Retention

Each Federal Agency is required by statute (36 CFR 1228) to maintain a comprehensive records schedule. This comprehensive schedule is developed by combining the General Records Schedules (containing disposal authority for records common to several or all agencies), published by the National Archives and Records Administration, with EPA-specific schedule items or record series.

A Records Control Schedule (schedule) constitutes EPA's official policy for records and information retention and disposal. The schedule (see Appendix E for samples) provides mandatory instructions for what to do with records (and non-record materials) no longer needed for current Agency business. Records retention and disposal should occur at regular intervals in the normal course of business of the Agency.

Other benefits of using schedules are:

- Ensures that the important records are organized and maintained in such a way as to be easily retrieved and identifiable as evidence of the programs activities (especially in the event of an audit, a FOIA request, or a discovery for a lawsuit).

- Conserves office space and equipment by using filing cabinets to house only active records.

- Saves money by the regular transfer of inactive files to less costly Federal Record Center (FRC) storage areas for subsequent disposition.

- Helps preserve those records that are valuable for historical or research purposes.

- Stabilizes the growth of records in offices through systematic disposition of unneeded records.

Based on careful analysis of the Agency's documentary materials, the schedules provide instructions for the retention and disposition of each record series or system

and of non-record materials, and authorizes the systematic removal of unneeded records from offices.[40]

Retention of records should not become an ad hoc process. A formal process is necessary for many reasons, including knowing the location of data, the format of data, and avoiding the need to keep data that are no longer useful or necessary. In an ad hoc process, data are kept in many different locations and systems and often on an individual rather than organizational basis, data formats will vary greatly from simple ASCII text files to specialized application formats, and the data may or may not be kept for the correct retention period.

EPA's NRMP provides the following process for record retention:

1. Take An Inventory
2. Appraise All Records
3. Establish Retention Periods
4. Develop Retention Schedule
5. Approve Retention Schedules
6. Apply Retention Schedules

Steps 1 and 2 of this process are the most time-consuming and tedious but also very important. Categorizing records during appraisal facilitates establishing retention periods and the retention schedule.

If data are already in electronic format, then they can be stored internally in the same format. If the data are to be archived at NARA, then it may have to be converted to paper format. At this point, retention internally of the electronic format will have to be decided until NARA updates their policy on electronic records.

NARA recognizes the need for an improved electronic records policy. In January 1998, they issued guidelines for digitizing archival materials but prefaced the guidelines with a statement that said: "The Guidelines do not constitute, in any way, guidance to Federal agencies on records creation or transfer to the National Archives of the United States."

[40] http://www.epa.gov/records/what/quest6.htm

Recognizing the need for a policy, NARA has instituted a "Fast Track Guidance Development Project" to identify currently available "best practices" and provide guidance quickly on electronic records issues that urgently confront Federal record keepers now. The intention is that "Fast Track Guidance" can be used while work goes forward on developing more complete and longer-term solutions.

The Target Audience for this guidance is agency Records Officers and records management staff and the intent is to aid their work with agency CIOs and IT staff on electronic recordkeeping and electronic records issues.

8.6 Ready and Reliable Access

Ready and reliable access to electronic information means that electronic access is available in a timely fashion, and that this access continues to be available in the future.

Some storage systems may provide electronic access but the methodology for access is not easily accomplished or possibly not reliable. For example, if file retrieval depends upon a modem connection, this may not be reliable. If file retrieval depends upon setting up gateways through different servers or through the mainframe, then the access is not readily available.

Access to data must include the ability to read or present the data in a format understandable to the user. For instance, if a magnetic tape were accessed to retrieve a data set, would the following conditions be satisfied:

1. Would the current computer hardware and version of DBMS be able to read the data set that was available when the application was archived?

2. Would the Operating System available today allow the DBMS application with which the dataset was created to run?

3. Would the storage device available today be compatible with the storage device needed by the DBMS application to access the data (i.e., does the application make device-specific commands when processing the data set)?

The opinion of many archivists regarding future access is illustrated by the following quote:

"Given the fact that digital recording technologies can be supplanted by a newer format every 5 to 10 years, the bigger problem facing archivists is the lifetime of the technology, not the lifetime of the medium...." [41]

[41] Dr. John W. C. Van Bogart, National Media Laboratory, 3M, St. Paul, Minn.

9 RECORDS MANAGEMENT SOLUTIONS

9.1 Records Management Software (DoD5015.2 Compliant)

This section summaries leading RMA software certified by DISA as compliant with the DoD 5015.2 specification. An evaluation of the software will require a separate study at a future date.

9.1.1 IBM RecordsManager

RecordsManager, the IBM EDM Solution for Records Management, delivers enterprise-wide document and records management in a single package. The system is built on the functions in Documentum, IBM's document management product. RecordsManager assures that all documents, whether hard-copy or electronic, move from creation to final disposition in a controlled environment. As each record matures, the system applies increasingly restrictive records management rules and assure that documents are not maintained any longer than absolutely necessary.

9.1.2 PSSoftware Solutions Ltds. - RIMS Studio 7.1

(http://www.pssoft.com/)

RIMS (Recorded Information Management Studio) 7.1 is a unified product suite that provides full life cycle management of all corporate records and information holdings, in paper or electronic format. It extends records management to other information management applications with integrated solutions which include email (allows users to file email records with a simple drag-and-drop technique) and electronic document management systems.

9.1.3 Filenet Corporation's Panagon IDMS 4.3 with RIMS 7.1.4

(http://jitc.fhu.disa.mil/recmgt/filenet/fn_sum3.htm)

Panagon/RIMS is an integrated product that combines the electronic document management capabilities of Panagon with the records management capabilities of

RIMS (see 9.1.2). The result is a system that stores and manages both official records and their predecessor "draft" documents to provide appropriate control over the life cycle of the items, from origination to disposition.

All electronic documents are stored in the Panagon repository. The RIMS software performs the bulk of the remaining functions required to schedule, screen, and dispose of the stored records in accordance with DoD 5015.2-STD.

9.1.4 Tower Software Corporation's TRIM 4.3

(http://www.ustrim.com/trim/index.htm)

TRIM is an off-the-shelf solution that enables the management of an organization's business information. TRIM combines electronic document management (EDM) functionality with information management principles to ensure an organization's business records are captured and made accessible in a secure, scalable, and manageable environment.

9.1.5 Provenance Systems Inc.'s ForeMost

(http://www.provsys.com/menu.html)

ForeMost is a suite of products for lifecycle management of electronic and paper records. ForeMost incorporates a secure repository for electronic records, a corporate filing system, formal retention and disposition scheduling, and other advanced administrative records control. End user services such as document filing, classification, and retrieval are supported. ForeMost will capture and store electronic mail, and any other type of electronic document. An Internet client version of ForeMost, called Odyssey, features true browser-based thin client records management. A Developer's Toolkit allows ForeMost to be integrated within desktop applications, for customer-designed records capture, classification, and retrieval.

ForeMost was the first records management software to be certified compliant with the US DoD 5015.2 standard for electronic records. Out of 14 products that have been registered compliant by the Joint Interoperability Test Command (JITC) of the Defense Information Systems Agency (DISA), over half are either ForeMost products or third party solutions integrated with ForeMost.

9.1.6 Dynosolutions Inc.'s CS-CIMS 2.5 with FOREMOST 6.3

(http://www.dynsolutions.com/d_rma.htm)

This product is now called KnowledgeFlow RMA. KnowledgeFlow RMA simplifies the challenge of capturing all the right documents, and it is certified as Year 2000 compliant and certified as meeting the DoD Standard 5015.2 for electronic records management. KnowledgeFlow RMA has tools and techniques to help users move their documents across the imaginary line of responsibility to become an official corporate record. KnowledgeFlow RMA's lookup feature provides document classification tools that give users techniques for fast, simple, and accurate classification. KnowledgeFlow RMA applies a three-stage records retention and disposition process. It automatically tags records qualified for destruction or archiving. The records manager can review the results with users prior to taking final action.

9.1.7 Filenet Corporation's Panagon IDM 4.2 with Foremost 6.3

(http://jitc.fhu.disa.mil/recmgt/docsopen/do_sum.htm)

Panagon IDM Desktop facilitates access and management of a variety of document types, including scanned paper documents, fax, word processing, spreadsheets, HTML forms, audio and video clips, computer-generated reports, and electronic document interchange (EDI) information. You can view more than 200 different document types without the native applications present. The ForeMost software is used in conjunction with IDM to perform the remaining functions required to schedule, screen, and dispose of the stored records in accordance with DoD 5015.2-STD.

9.1.8 PCDOCS Inc.'s DOCS Open 3.7.2 with Foremost 6.3

(http://jitc.fhu.disa.mil/recmgt/idmds/rimssum.htm)

DOCS Open provides electronic document management capabilities, and includes its own document repository. The ForeMost software provides file plan implementation, scheduling, screening, and disposition of the records stored in the DOCS Open repository. A gateway integrates the ForeMost records management application with DOCS Open. Documents in the DOCS Open repository can be "synchronized" with ForeMost via the gateway. Synchronizing a document creates a corresponding

document profile in ForeMost and restricts access to the original document in the repository. This procedure places the document under formal records management control. DOCS Open and ForeMost together satisfy all the mandatory requirements of DoD 5015.2-STD.

9.2 Records Management Software (Other)

Although this study recommends the purchase of DoD 5015.2-compliant software, there are other RMA products that are worth considering because of synergies with EPA's IT architecture and because they may currently be in the process of obtaining DoD 5015.2 compliance certification.

For instance, Lotus Development has products in their Domino family that manage and archive records. Because EPA's enterprise mail system is being standardized on Lotus Notes Mail, and since email may be a very important consideration in records management, the Lotus product should be considered. This section summarizes two currently non-compliant RMA products.

9.2.1 Lotus Domino.Doc Records Management (LRM)

Domino.Doc Records Management is an add-on for Domino.Doc, Lotus' document management system. It is an integrated Domino.Doc add-on that is an automated solution for the retention and disposition of electronic records based on a customer's unique corporate policies. It leverages features of Domino.Doc such as storage repository, file cabinets, binders, templates, menu items, and the user interface. It also leverages Lotus Domino's administration, replication, messaging, and security models.

9.2.2 Extemporé Information Management and Tracking Software

Extemporé is designed and written exclusively for Windows. It identifies, classifies, traces, and manages all forms of information from inception through destruction or archive. The technologies used in this system include Inter/Intranet web browser-based search and request capabilities, OLE 2.0/ActiveX technology for electronic document and image linking, HTML (HyperText Markup Language)/Web publishing, MAPI and VIM electronic mail integration and a client/server architecture employing SQL (Structured Query Language) queries and ODBC (open database connectivity) drivers.

10 STORAGE SOLUTION CRITERIA AND EVALUATION

The criteria for a storage solution discussed in this section pertain primarily to storage devices. Storage architectures (such as SAN) and storage management criteria and evaluation require further investigation of those technologies and of EPA's requirements.

10.1 Criteria for Storage Systems

Seven criteria identified for storage systems selection are described below:

10.1.1 Access Time

Access time is the time a user has to wait to receive information from a storage system. Storage systems use many different measures that are part of access time, such as seek time, read time, and transfer time. Seek time is the time it takes to locate the data on the device and is generally lowest for hard disk storage and highest for tape storage. Read time is the time it takes to read a sector of data once the read head is positioned over the data with tape storage generally having the lowest time and optical disc the highest. Transfer time is the time it takes to move information from the media through the electronics and connections to the computer's memory. This time is a function of the quality and parallelism of the electronics and the type of connection used from the storage device to the computer.

10.1.2 Capacity

Capacity is the maximum amount of data that can be stored on a storage device. This is a quantity that varies greatly based on the media and the configuration of the media. In general, tape storage has the highest capacity followed in decreasing order by magneto-optical, magnetic disc, and optical disc.

10.1.3 Cost

Storage systems have a very wide cost range independent of the media. However, cost comparison between storage system can be done by dividing the total cost of the

system and media by the capacity. The units for this measure are normally dollars per megabyte ($/Mbyte). When rated in $/Mbyte, tape systems are the least expensive followed in increasing cost by DVD, magnetic disc, magneto-optical disk, and CD.

10.1.4 Availability

Availability refers to how easy it is to access data. If data are stored on-line, availability is very good. If data are stored off-line, and retrieval is possible by commands from the user's terminal, then availability is good to fair. If data is stored off-line, and retrieval involves filling out forms and waiting for approval or processing, then availability would be low. Type of media does affect availability. For instance, magnetic disc storage generally has very high availability because it is not generally removable, while magnetic tape might have the lowest availability (if it is stored offline).

10.1.5 Longevity

Longevity or Shelf-Life of media is the amount of time that a medium will preserve the data that is stored. Paper has one of the highest longevity ratings, and magnetic disc or tape the lowest. Longevity is affected by the quality of the media and by the storage environment. In general, high quality media last significantly longer whether the medium is paper or more recently optical discs. The storage environment for most media should be free of extreme temperatures swings and have low humidity. Temperature should be normal room temperature or slightly cooler (~70 F).

A second aspect of longevity is the technology of the storage system. If media have sufficiently long shelf life to last 50 years, the question then becomes whether or not the storage systems to read the media will be available as well, and in what state of maintenance.

10.1.6 Security

Security of media depends on the type of storage system and on the type of storage media. Systems that require manual intervention (such as a tape drive) carry inherently higher security risk than systems that handle the media automatically, and can be

secured more effectively. Media that are re-writable are less secure than write-once media since the latter does not allow the data to change.

10.1.7 Portability

Portability refers not only to the ability to physically remove and transport the media, but also to cross-platform compatibility. Optical technologies such as CD or most DVD use non-proprietary read/write formats and have the greatest portability characteristics. A CD made on any operating system using any CD recorder can be accessed from any computer using any CD-ROM reader.

10.2 Evaluation of Storage Systems Based On Criteria

Storage systems must be carefully evaluated to be sure they meet the criteria of the application they are intended for. If access time is the most important criteria, then the system will most likely use magnetic disc. If the cost of storing large amounts of data is important, then the choice will likely be magnetic tape or DVD optical disc. In general, access time, capacity, and cost are the most important tradeoffs that need to be considered in selecting a storage solution.

10.2.1 Storage of Active Records

For storage of active records (frequently used and infrequently used), we rank the seven criteria in the following order of importance from most important (1) to least important (7):

1. Availability
2. Access time
3. Capacity
4. Security
5. Cost
6. Portability
7. Longevity

Based on this ranking, the recommended solution for storage of active records would be high-performance magnetic disk for frequently used records integrated with a nearline solution employing automated tape libraries or optical disk jukeboxes for infrequently used records.

10.2.2 Internal Archival of Records

For the internal retention of records at EPA, we rank the seven criteria in the following order of importance from most important (1) to least important (7):

1. Capacity
2. Cost
3. Longevity
4. Portability
5. Security
6. Availability
7. Access time

Based on this ranking, the recommended storage solution for EPA's internal data archives would be a system employing magnetic tape or DVD disc and magnetic disk. Both magnetic tape and DVD disc can store large quantities of data at relatively low $/Mbytes.

10.2.3 External Archival of Records

External archiving of records presumes that EPA will transfer records to a FRC operated by NARA. Currently, NARA is evaluating new guidelines for electronic records submitals. Their pre-existing guidelines call for 7 or 9-track tape and data formatted in ASCII or EBCDIC. Obviously, this guideline is too antiquated, so NARA may accept newer media and formats although official policy has not been announced.

Paper medium still accounts for the majority of data submitted by EPA. NARA, in collaboration with DoD (see Section 6.5), is deciding on a preferred format for electronic imaging of paper and of electronic files.

In terms of media, all the government agencies recently surveyed (see Section 6.5) are already archiving to optical media. NARA is not averse to this choice, but is concerned with receiving optical discs in too many different varieties (see Section 6.2.1), some of which may soon be outdated technology.

For delivery of electronic records to NARA, we rank the seven storage criteria in the following order of importance from most important (1) to least important (7):

1. Portability
2. Longevity
3. Security
4. Cost
5. Capacity
6. Availability
7. Access time

Based on this ranking, the recommended storage solution for EPA's external electronic data archives would be a storage system with DVD optical disc. The medium for electronic records with the highest longevity is currently high-quality optical disc (see Section 6.3). The highest capacity and the lowest cost is provided by magnetic tape first and optical disc second; however, optical disc is more durable under temperature and humidity extremes. Both are portable, but DVD is inherently more secure in a write-once format. Access time is much better with optical disk than magnetic tape because data is randomly accessed. Availability for both media is about the same, but magnetic tape drive technology is becoming less prevalent than optical disc drives,

and this trend is expected to continue. DVD optical disc is therefore suggested given the above criteria priority.

11 RECOMMENDATIONS

The following sections address recommendations for storage media, RMA software, and an overall records management storage and storage management architecture.

11.1 Storage Solutions

As EPA moves forward with electronic records storage and management, the choice of storage media will become critical. If access time is the most important criterion, then the system will most likely use magnetic disc. If the cost of storing large amounts of data is important, then the choice will likely be magnetic tape or DVD optical disc. In general, access time, capacity, and cost are the most important tradeoffs that need to be considered in selecting a storage solution.

The most prevalent approach for combining these objects is through the implementation of a hierarchical storage management (HSM) methodology using the concepts of on-line, near-line, and off-line records. Under such a strategy, the most cost-effective solution can be selected based upon access requirements, cost constraints, security, and capacity needs. Magnetic disk is still the preferred choice for meeting frequent on-line access requirements. Based on the market offerings available today, and the forecasted trends, optical media or magnetic tape appears to be the best choice for near-line and off-line storage. The highest capacity and the lowest cost is provided by magnetic tape first and optical disc second; however, optical disc is more durable under temperature and humidity extremes. Access time is much better with optical disk than magnetic tape because data is randomly accessed. Availability for both media is about the same, but magnetic tape drive technology is becoming less prevalent than optical disc drives, and this trend is expected to continue. DVD optical disc is the suggested choice given the above constraints, however, further study is required to select an optical format along with the hardware and software necessary to support records storage requirements at each regional location.

11.2 RMA Software

There are many applications available today to manage the workflow and disposition of records information. Some packages integrate better than others with the systems already in place in an organization. Some packages have greater capability in workflow control, while others are better suited for managing storage.

On April 11, 1997, DoD reissued directive 5015.2 (see Appendix C) which defines a list of requirements for records management. The directive has been accepted by NARA as guidance in this area. The directive requires the Defense Information Systems Agency (DISA) to establish and maintain the capability to test and evaluate records management applications that are seeking DoD approval and lists criteria in broad terms. As a result of DoD's involvement and NARA's acceptance, this directive has become an ad hoc standard for Records Management Application (RMA) software (see Section 9).

As of September 10, 1999, the following Records Management Applications[42] have been approved by the DISA, Joint Interoperability/Test Command as being compliant with the DoD5015.2 standard (see Appendix C for a description of the standard):

- IBM's RecordsManager 1.1
- PSSoftware Solutions Ltd.'s RIMS Studio 7.1
- Filenet Corporation's Panagon IDMDS 4.3 with RIMS 7.1.4
- Tower Software Corporation's Trim 4.2
- Tower Software Corporation's Trim 4.3
- Provenance Systems Inc.'s Foremost 6.3
- Provenance Systems Inc.'s Foremost 7.0
- Provenance Systems Inc.'s Foremost Enterprise 1.0
- Dynsolutions Inc.'s CS-CIMS 2.5 with Foremost 6.3
- Filenet Corporation's Panagon IDM 4.2 with Foremost 6.3
- PCDOCS Inc.'s Docs Open 3.7.2 with Foremost 6.3
- Eastman Software's DMX 1.1 with Foremost 7.0
- Universal Systems' Inc.'s E.POWER 1.5
- PCDOCS Inc.'s Docs Rm 3.1

[42] http://jitc.fhu.disa.mil/recmgt/

Selection of an RMA must involve a careful assessment of business processes, system integration, records flow, records volume, and management objectives. Such an assessment is beyond the scope of this paper, but must be conducted before selecting an RMA product. As EPA comes closer to beginning an electronic records management program, further study will need to be done to select software that will meet the requirements of the DoD directive, NARA, and the EPA Records Management Program office.

11.3 Architecture Alternatives

The emergence of low cost multiprocessor hardware architectures, coupled with development of server-based operating systems and database technologies, has allowed organizations to transfer their most valued assets from large centralized computing platforms to distributed LAN-based architectures. This trend has resulted in dramatic improvements in PC-based architectures in the areas of operating systems, security, network capabilities, and storage peripherals and subsystems. Based on current hardware and software technologies, it is feasible to consider two primary architectural alternatives for electronic records management:

- **Centralized**. All active (on-line), relatively inactive (near-line), and inactive or archived (off-line) records are stored at the National Computer Center (NCC). In such an architecture, once the responsible information owner declares the information to be a record, the end-user application immediately transfers an official record copy to the central facility for safekeeping for the life of the record, in accordance with the appropriate Agency retention schedule. In addition, the Records Management Application (RMA) creates centrally stored metadata for the record. The central facility is also the repository for all electronic records that enter the Agency through external sources. The central RMA will schedule, as required, transfer of records to the Federal Records Center for archiving by NARA.

- **Distributed**. All active (on-line), relatively inactive (near-line), and inactive or archived (off-line) records are stored at the local EPA regional data centers (RDC). In such an architecture, once the responsible information owner declares the information to be a record, it is transferred to the RDC facility for safekeeping for the life of the record, in accordance with the appropriate Agency retention

schedule. The regional data center would also be the repository for all electronic and paper records that enter the Agency through external sources. The RDC's RMA will create and store metadata relating to each record in the RDC. The RDC's RMA facility will then schedule the transfer, as required, to the regional Federal Records Center, of those records that are to be archived by the FRC and NARA.

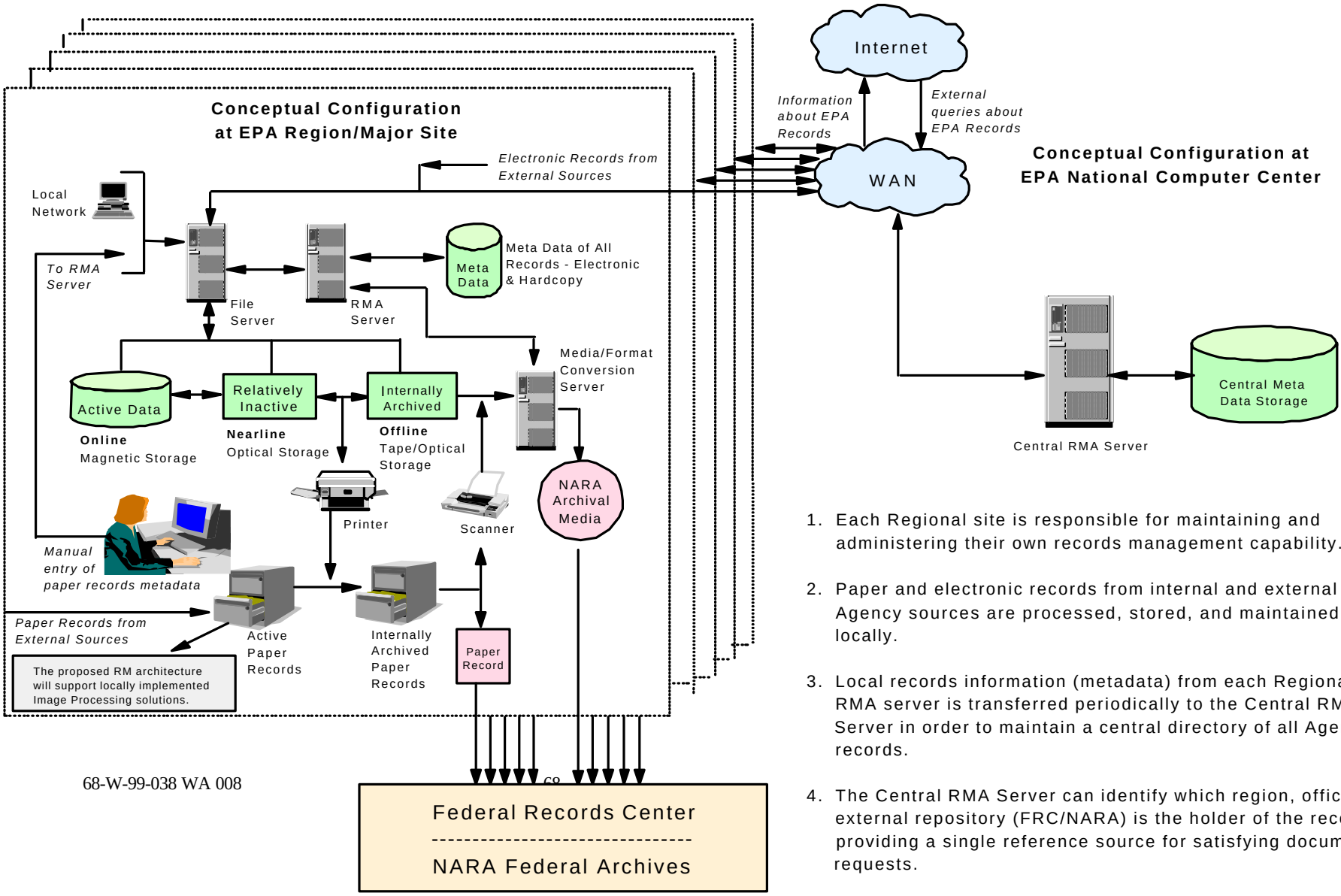
The SRA Team recommends an approach very closely aligned with a fully distributed architecture. Such an approach makes each regional office responsible for the proper administration, maintenance and support of their records program. Due to the large volume of records created in the Agency, and the often highly sensitive nature of their content, it is most practical to keep the information as close to the record custodian as possible.

Another consideration concerns cost, and the applicable allocation of operational expenses. With a central storage approach, a charge-back methodology that ensures each region pays for their share of the central infrastructure burden is required. While feasible, a central approach would also force EPA regions to abide by a "one size fits all" solution. A distributed approach, tailored to regional needs and installed infrastructure components, is likely to provide the most acceptable solution, giving the greatest amount of control to the regions, in the most equitable, cost-effective, and scalable manner.

11.4 Details of Distributed Architecture Alternative

While each region will be responsible for the management and storage of their own records, the system will use a central metadata repository. The repository, to be located at the EPA National Computer Center (NCC), will provide a single reference source for information regarding the nature, disposition, and location of all records held by the Agency. Figure 11-1 provides a conceptual view of the recommended architecture components for both a Regional/Major site configuration, and the NCC.

Figure 11-1. Conceptual View of Recommended Architecture Components



1. Each Regional site is responsible for maintaining and administering their own records management capability.

2. Paper and electronic records from internal and external Agency sources are processed, stored, and maintained locally.

3. Local records information (metadata) from each Regional RMA server is transferred periodically to the Central RMA Server in order to maintain a central directory of all Agency records.

4. The Central RMA Server can identify which region, office, or external repository (FRC/NARA) is the holder of the record, providing a single reference source for satisfying document requests.

11.4.1 Conceptual Records Management Architecture at an EPA Region/Major Site

EPA generates records from many sources in operational and administrative areas. The operational sources include the scientific collection and analysis of data, regulatory requirements and directives, and the published and unpublished results of studies. Administrative sources include email, human resource information, financial data, policies and procedures information, and internal directives.

Records may originate in either paper or electronic format. The goal of the selected architecture is to provide flexibility to each region to decide how they wish to store records data once they are received. Electronic documents can be stored electronically or printed on paper. Paper documents can be stored as paper records, or imaged and processed electronically.

Regional users who normally process records data will be able to do so via a local area network connected to their Regional data center. The users' desktop applications (e.g., e-mail, word processing) will be augmented with the capability (provided by the records management application, or RMA) of designating a particular document they are processing as an official record. As part of the automated process, the user will indicate the necessary records schedule information which includes information such as record custodian, location, description, retention rules, and so forth. This data will form the basis of the metadata record that will be used to track the record throughout its lifecycle. The record, along with the metadata tag, is then transferred to the regional data center.

Once the record and metadata files are transferred to the RDC, the record itself is stored on on-line magnetic media attached to a file server. The metadata record is transferred to the RMA server. On a periodic basis, metadata records from the RDC are transferred to the NCC for backup and centralized records tracking purposes. Backup of the records at the RDC are governed by EPA policy for general support computer facilities.

The record will remain in the custodial care of the RDC until it is either destroyed or transferred to a Federal Records Center under the auspices of NARA. The management of the record within the RDC will be governed by the records schedule and the principles of hierarchical storage management (HSM). HSM uses the concept

of on-line, near-line, and off-line storage facilities. On-line storage is recommended to be magnetic disk because of its fast access time, flexibility, and high availability. However, because of its relatively high cost, records residing on on-line storage are typically migrated to near-line following a pre-defined period of inactivity.

Similar to on-line storage, near-line storage is still accessible without operator intervention, but typically with greater access times. Recommended near-line media such as magnetic tape or DVD have longer access times, but at significantly less cost than magnetic disk.

As the name implies, off-line storage is not available without manual intervention. Magnetic tape is the recommended medium for this capability. While magnetic tape has a longer access time than disk due to its sequential (non-random) format, magnetic tape has the lowest price point of any currently available storage media. Additionally, the anticipated minimal access requirements for records that have migrated to off-line storage makes the longer access time insignificant.

Initially, it is anticipated that most records that enter the Agency in paper format will be retained as paper. Those records that enter electronically will be maintained as electronic records. However, the ability to image a paper record and store and maintain it electronically will be provided. The existing SCRIPS image processing capability is an example of an in-house paper-to-electronic records transformation process that will need to be integrated and supported by the new records management architecture. Regardless of how a paper record is to be transformed or stored, the necessary retention schedule information and required RMA metadata must be manually entered into the RMA system for tracking purposes.

As indicated by retention schedules, certain records will remain archived in-house at EPA until they are required to be destroyed. Other records will be transferred to a Federal Records Center and remain archived under NARA custodianship until their destruction date. In either case, the RMA system will track the record location and disposition throughout the record life cycle.

Part of the process of archiving digital data is converting existing digital data to a digital archival format, and converting data in paper or microfilm form to that same digital archival format. A consistent well-supported format becomes critical to information accessibility and usability. While it is possible to store digital data in its

native format (i.e., a WordPerfect 8 file, or a Freelance Graphics file) this creates problems with discontinued software, and possibly problems as newer versions of software are published. This study recommends the use of Tagged Image File Format (TIFF) and Portable Document Format (PDF). TIFF and PDF are the two most common image formats in use by the government today. EPA will need to further evaluate to determine if these formats comply with Agency and NARA requirements.

11.4.2 Conceptual Records Management Architecture at the NCC

The purpose of the central RMA metadata repository at the NCC is to provide a central collection point for information pertaining to all EPA records. Access to this repository will be available internally to EPA via the wide area network. Access by external entities will be available via the Internet. To avoid the possibility for unauthorized disclosure of records data, direct access to actual records will not be available via the Internet; however, point-of-contact information can be provided via the metadata record.

11.5 Recommended Future Actions

The selected records management architecture must meet Agency requirements for secure and reliable records storage, while providing a fast and efficient means for records storage and retrieval. To meet these objectives, the following key elements must come together in a synergistic fashion:

- Implementation of a robust, standards-compliant records management application software system. It is recommended that the selected RMA software be compliant with DOD Directive 5015.2-STD, Design Standard for Electronic Records Management Software Applications.

- Implementation of a process that will accommodate records tracking regardless of the format in which they enter the Agency (i.e., electronic or paper records).

- Implementation of a process that will allow for the transfer of records to NARA in NARA-approved media types.

- Selection of portable, reliable, cost-effective storage media that meet storage life requirements mandated by records retention schedules.

- Establishment of a cataloging process that allows for the automated management and tracking of electronic records from the time they enter the Agency, until they are either destroyed or delivered to NARA for continued safekeeping.

- Implementation of security measures that ensure classification, control, access, and release of records remains with the document custodian.

- Design of an architecture that integrates in a near-seamless fashion with approved and implemented Agency hardware, software and network technologies.

In addition to the findings documented in this paper, the following additional analysis is recommended:

- A study is needed of the 600 plus Record Control Schedules to determine what percentage of data is sent to NARA and what percentage of data is retained by EPA and not sent to NARA. This would provide for a better estimate of the volume of data that EPA is retaining internally, and is necessary to properly size any storage system solution acquired for archival of data.

- Selection of an RMA must involve a careful assessment of business processes, system integration, records flow, records volume, and management objectives. As EPA comes closer to beginning an electronic records management program, further study will need to be done to select software that will meet these objectives in addition to the requirements of the DoD directive, NARA, and the EPA National Records Management Program office. Additionally, although this study recommends the purchase of DoD 5015.2-compliant software, there are other RMA products that are worth considering because of synergies with EPA's IT architecture and because they may currently be in the process of obtaining DoD's compliance.

- The alternatives for storage media accepted by NARA are paper and open reel tape. This may soon change to include cartridge tape and optical disk when NARA releases their guidance for electronic records submitals. EPA can, however, choose whichever storage medium that meets the cost/benefit and operational

needs of the Agency for internal records archival. DVD optical disc is the overall archival media of choice; however, further study is required to select an optical format along with the hardware and software necessary to support records management requirements at each regional location.

C EPA will need to select an image format for electronic storage of records. TIFF and PDF are front runners in this area today, and conform to recommendations made by DoD. EPA will need to further evaluate to determine if these formats comply with Agency and NARA requirements.

The implementation of an electronics records management process is a significant undertaking, and with the rapid changes in technology, an undertaking that will require constant and continual attention. Even the selection of a storage medium and storage system may not be longlasting:

> "Given the fact that digital recording technologies can be supplanted by a newer format every 5 to 10 years, the bigger problem facing archivists is the lifetime of the technology, not the lifetime of the medium...."[43]

If this perception is a reasonable assumption given the fast pace of technology change, then EPA must be prepared to migrate their archive storage system to newer technology every 5 to 10 years at a cost of 50 to 100% of the initial hardware investment.

In general the advantages of electronic records management outweigh the disadvantages of the manual processes for tracking and storing paper records. With the sheer volume of data that is being earmarked for archival, most federal Agencies see an automated electronic storage method as the only effective means of dealing with the profusion of data caused by the increasing use of information processing systems.

[43] Dr. John W. C. Van Bogart, National Media Laboratory, 3M, St. Paul, Minn.

APPENDIX A. GLOSSARY OF TERMS

Access Time	The time it takes to retrieve a piece of information. With hard disks or compact discs, maximum access time is measured as the time it takes to move from one end of the disk to the other, find a piece of information, and transfer that information to RAM.
BLER	BLock Error Rate. This is the “raw” digital error rate before any error correction, which in turn is a determining factor in the quality and readability of a given disk..
BLERmax	The maximum number of BLERs allowed on a disc. According to the industry standard, a CD-ROM is allowed a BLER of up to 220 before it is considered a "bad" disc.
bubble memory	A type of non-volatile memory composed of a thin layer of material that can be easily magnetized in only one direction. When a magnetic field is applied to circular area of this substance that is not magnetized in the same direction, the area is reduced to a smaller circle, or bubble.
buffer underrun	The most common problem in CD recording. An underrun occurs when the system cannot keep up a steady stream of data as required by CD recording. The CD recorder has a buffer to protect against interruptions and slowdowns, but if the interruption is long enough that the recorder's buffer is completely emptied, a buffer underrun occurs, writing halts, and most often the recordable CD is irretrievably damaged.
buffer	An amount of memory which temporarily stores data to help compensate for differences in the transfer rate from one device to another.
byte	One character, made up of 8 bits.

cache	A portion of RAM used for temporary storage of data which must be accessed very quickly. In applications which run from CD-ROMs, the cache is typically used to store directory files.
capacity	The maximum amount of data that can be stored on a storage device.
CBI	Confidential Business Information. Information submitted by private industry that must be safeguarded.
CCD	Charge-Coupled Device. An instrument whose semiconductors are connected so that the output of one serves as the input of the next. Digital cameras, video cameras, and optical scanners all use CCD arrays.
CCITT	Consultative Committee on International Telephone and Telegraphy. Now known as the ITU-T (for Telecommunications Sector of the International Telecommunications Union), it is the primary international body for fostering cooperation standards for telecommunications equipment and systems. This international committee recommends telecommunication standards including the audio compression/decompression standards.
CD/DVD-ROM	A new format jointly developed and agreed upon by Toshiba, Matsushita, Sony, Phillips, Time Warner, Pioneer, JVC, Hitachi, and Mitsubishi Electronics. CD/DVD-ROM is an acronym for Compact Disc/Digital Versatile Disc, which is now the universal format for high density compact discs. For computer applications, the disc will be called DVD-ROM and for audio, DVD-Audio. The new format has the capacity of holding more information, and has a higher data transfer rate.
CD	Compact Disc. A digital medium formed of a 12cm polycarbonate substrate, a reflective metalized layer (aluminum, gold, or silver), and a protective lacquer coating. The physical format of CDs is described by the ISO9660 industry standard.
CD-WO	Compact disc-Write Once. Recordable compact disc.

CD-R	Compact Disc-Recordable. This term is used to describe the technology of recordable CD as well as the equipment, software and media used to make recordable discs. CD-Recordable discs have an organic dye layer between the substrate and the metal reflective layer that serves as the data image. Laser pulses change the transparency of the dye according to the data bits.
CD-RW	Compact Disc-ReWritable. The most recent addition to the compact disc family. It was originally called "CD-Erasable." It is a media and recording system that allows the user to erase previously recorded information and then record new information onto the same physical location on the disk.
Cross-talk	This is a measure of the amount of interference coming from neighboring pit tracks on a CD. As track pitch is tightened (when tracks are packed closer together to put more data on a disc), cross-talk increases. A maximum value of 50% is allowed by Red Book specifications.
Cyanine	One type of organic dye used to form the data layer in CD-R discs. Cyanine was the first material used for these discs, but presently a metal-stabilized cyanine compound is generally used instead of “raw” cyanine. An alternative material is phthalocyanine.
DAT	Digital Audio Tape. A type of magnetic tape that uses a scheme called helical scan to record data. A DAT cartridge is slightly larger than a credit card in width and height and contains a magnetic tape that can hold from 2 to 24 gigabytes of data. It can support data transfer rates of about 2 MB/s. Like other types of tapes, DATs are sequential-access media.

data layer	In CD-R, the organic dye sandwiched between the polycarbonate substrate and the metalized reflective layer of the media. CD-Recordable discs do not have any data on them until they are recorded. Instead the recording laser selectively melts "pits" into the dye layer—but rather than burning holes in the dye, it simply melts it slightly, causing it to become non-translucent so the reading laser beam is refracted rather than reflected back to the reader's sensors. In pressed CDs, the data layer is part of the polycarbonate substrate, and is pressed into the top side of it by a "stamper" during the injection molding process.
DLT	Digital Linear Tape. A type of magnetic tape storage device originally developed by DEC and now marketed by several companies. DLTs are ½-inch wide and the cartridges come in several sizes ranging from 20 to over 40 GB. DLT drives are faster than most other types of tape drives, achieving transfer rates of 2.5 MB/s.
DoD 5015.2	Department of Defense Records Management Application Directive (April 11, 1997)
ECC	Error Correction Code. A system of scrambling data and recording redundant data onto disk as it is premastered. On playback, this redundant information helps to detect and correct errors that may arise during data transmission.
EDC	Error Detection Code. Thirty-two (32) bits in each sector which are used to detect errors in the sector data.
EDMS	Electronic Data Management System.
ERM	Electronic Records Management. Term representing the handling of digital data.

fault tolerance	The ability of a system to respond gracefully to an unexpected hardware or software failure. There are many levels of fault tolerance, the lowest being the ability to continue operation in the event of a power failure. Many fault-tolerant computer systems mirror all operations—that is, every operation is performed on two or more duplicate systems, so if one fails the other can take over.
fibre channel	A new technology for communicating with storage devices that can be used to replace the SCSI interface between servers and storage.
file system	A data structure that translates the physical (sector) view of a disk into a logical (files, directories) structure, which the application and user can more easily use to locate files.
FOIA	Freedom Of Information Act. Act of the U.S. Government requiring agencies to release information requested by the public citizen.
FRC(s)	Federal Records Center(s). Regional branches of NARA that accept records for archival storage.
Gigabyte (GB)	One gigabyte (1GB) is equivalent to 1,024 Megabytes of data.
HSM	Hierarchical Storage Management. A system that integrates the three types of storage so that less-accessed data is placed on less-costly media.
Indexing	Creation of a data index to speed up search and retrieval.
Jukebox	A robotic system for accessing tapes or CDs where the media/drive ratio is > 1. May include a hard disk cache. Typically includes software for optimizing access.
Kilobyte (KB)	1 Kilobyte is equivalent to 1,024 bytes of data.
lacquer spincoat	Acrylic lacquer is spincoated in a thin layer on top of the metal reflective layer of a CD to protect it from abrasion and corrosion. Usually a decorative label is also applied on top of the lacquer, but this is not a standard requirement.

logical block	The smallest addressable space on a disk. Each logical block is identified by a unique Logical Block Number (LBN), assigned in order starting from 0 (zero) at the beginning of the disk.
LZW	Lempel-Ziv Welch. A compression algorithm which provides a way of compressing data (primarily raster data) that takes advantage of repetition strings in the data.
magnetic tape	A magnetically coated strip of plastic on which data can be encoded. Storing data on tapes is considerably cheaper than storing data on disks. Tapes also have large storage capacities, ranging from a few hundred kilobytes to several gigabytes. Accessing data on tapes, however, is much slower than accessing data on disks. Tapes are sequential-access media, which means that to get to a particular point on the tape, the tape must go through all the preceding points. In contrast, disks are random-access media because a disk drive can access any point at random without passing through intervening points. Because tapes are so slow, they are generally used only for long-term storage and backup. Data to be used regularly is almost always kept on a disk. Tapes are also used for transporting large amounts of data.
magnetic disk	A round plate on which data can be encoded. Data is encoded as microscopic magnetized needles on the disk's surface. You can record and erase data on a magnetic disk any number of times, just as you can with a cassette tape. Magnetic disks come in a number of different forms: floppy disk, hard disk, removable cartridge.
mastering	Mastering is the process of creating a stamper or set of stampers to be used in the injection molding stage of manufacturing compact discs.

mastering	Technically, refers to the process of creating a glass master from which compact discs will be reproduced in quantity. In desktop recordable CD systems, mastering is done together with premastering by the desktop CD recorder, and the term is generally used to mean "recording."
Megabyte (MB)	1 Megabyte is equivalent to 1,024 kilobytes of data.
mirroring	A technique in which data is written to two duplicate disks simultaneously. This way if one of the disk drives fails, the system can instantly switch to the other disk without any loss of data or service. Disk mirroring is used commonly in on-line database systems where it's critical that the data be accessible at all times.
NARA	National Archive and Records Administration. Federal organization tasked with preserving the informational heritage of the federal government.
near-line storage	A storage unit that is connected to a computer system through an automatic storage interface system. Typically has slower access times but high capacity. Example: StorageTek silos or CD Jukeboxes.
NRMP	National Records Management Program (EPA). Program created inside EPA to manage record information. Recognized and supported by NARA.
NSIC	National Storage Industry Consortium. Membership consists of over fifty corporations, universities and national labs with common interests in the field of digital information storage. Corporate membership includes most major U.S. storage product manufacturers and many other companies from the storage industry infrastructure. NSIC has its headquarters in San Diego and was incorporated in April 1991 as a non-profit mutual benefit corporation. (http://www.nsic.org/)

off-line storage	A storage unit that requires manual intervention to make available storage volumes to a computer system. Typically has slowest access time but virtually unlimited capacity. Example: Open reel magnetic tape library with tape drive.
on-line storage	A storage unit that is connected directly to the I/O bus of a computer system. Typically has fastest access time but less capacity. Example: Hard Disk.
optical disk	A storage medium from which data is read and to which it is written by lasers. Optical disks can store much more data—up to17 gigabytes (6 billion bytes)— than most portable magnetic media, such as floppies. Optical disks record data by burning microscopic holes in the surface of the disk with a laser. Optical disks come in four basic forms: CD-ROM, WORM, erasable optical (EO), and DVD.
Orange Book	The Orange Book is the specification for recordable CD.
organic dye	The data layer of CD-R discs is made from either cyanine or phthalocyanine dye which is melted by the laser beam during the recording process. Where the dye is melted, it becomes opaque or refractive, scattering the reading laser beam so it is not reflected back into the reader's sensors.
phthalocyanine	An organic dye used to form the data layer in some CD-Recordable discs. Mitsui Toatsu Corporation holds the patent on this dye, but has licensed its formula to some other manufacturers.
pits & lands	In a “pressed” or mass-replicated CD, the bumps and grooves that represent the binary data on a disc's substrate are pressed into the plastic substrate during manufacture. CD-R discs do not have substrate pits and lands.

premastering	The process of preparing data to be recorded onto a compact disc. This includes breaking the data into sectors and recording those sectors with the appropriate header (address) and error correction information. In the case of recordable CD systems, premastering and mastering are done in one operation, resulting in a ready-to-read compact disc.
PROM	Programmable Read-Only Memory. A PROM is a memory chip on which data can be written only once. Once a program has been written onto a PROM, it remains there forever. Unlike RAM, PROMs retain their contents when the computer is turned off. The difference between a PROM and a ROM (read-only memory) is that a PROM is manufactured as blank memory, whereas a ROM is programmed during the manufacturing process.
RAID	Redundant Array of Independent Disks. A redundant subsystem of disk drives that improves performance, fault tolerance, and data recovery.
RAIT	Redundant Array of Independent Tapes. A redundant subsystem of tape drives that improves performance, fault tolerance, and data recovery.
RAM	Random Access Memory. A type of computer memory that can be accessed randomly; that is, any byte of memory can be accessed without touching the preceding bytes. RAM is the most common type of memory found in computers and other devices, such as printers. In common usage, the term RAM is synonymous with main memory, the memory available to programs.
reflective layer	The metal layer on top of the dye that reflects the laser beam back to the reading assembly. This is usually 24K gold in CD-Recordable discs, but recently silver has been introduced for CD-R discs as well. In CD-ROMs, the layer is usually aluminum.
RM	Records Management. Term to describe the handling of records.

RMA	Records Management Application - term used to describe the category of software available in the marketplace to manage archival of information.
ROM	Read-Only Memory. Computer memory on which data has been prerecorded. Once data has been written onto a ROM chip, it cannot be removed and can only be read. Unlike main memory (RAM), ROM retains its contents even when the computer is turned off. ROM is referred to as being nonvolatile, whereas RAM is volatile.
SAN	Storage Area Network. A managed high-speed network that provides any-to-any interconnection of server and storage elements.
SCSI	Small Computer System Interface (pronounced “scuzzy”). A standard for high-speed data transfer between computers and their peripheral devices. A SCSI interface allows up to eight different peripheral devices to be connected to a single controller.
sector	The smallest recordable unit on a CD. A disk can contain [(75 sectors per second) x (60 seconds per minute) x (number of minutes on disk)] sectors. The amount of data contained in the sector depends on what physical format and mode it is recorded in; for "regular" CD-ROM (Mode 1) data, you can fit 2048 bytes (2 kilobytes) of data into a sector.
SNIA	Storage Networking Industry Association. The Storage Networking Industry Association is an international computer system industry forum of developers, integrators, and IT professionals who evolve and promote storage networking technology and solutions. (http://www.snia.org/)
stamper	The data-bearing removable "die" used during the injection molding of a CD to imprint pits and lands into the polycarbonate substrate of the disc. In manufacturing CD-R media, instead of pits and lands, a continuous spiral is pressed into the substrate as a guide to the recorder's laser.

striping	A technique for spreading data over multiple disk drives. Disk striping can speed up operations that retrieve data from disk storage. The computer system breaks a body of data into units and spreads these units the available disks. Systes that implement disk striping generally allow the user to select the data unit size or stripe width. Disk striping stores each data unit in only one place and does not offer protection from disk failure.
substrate	The optical-quality injection molded clear polycarbonate plastic "bottom" of a CD or CD-R. For CD-Rs, this layer does not contain "pits and lands" but has a single spiral groove that guides the recorder's laser.
transfer rate	The amount of data which is transferred from the CD-ROM to the computer. The CD-ROM transfer rate is limited by the speed at which the disc rotates in the drive. The conventional CD-Rom transfer rate is approximately 150 kilobytes/sec, referred to as 1x. Therefore, a quadruple speed (4x) CD-ROM drive can transfer data at a rate of 600 KB/Sec.
WNRC	Washington National Records Center (FRC located in Suitland, Maryland)
WORM	Write Once Read Many - an acronym for media that can be written to once but are not erasable. CD-Recordable is an example.
Yellow Book	The book which sets out the standard for the physical format of compact discs to be used for information storage (CD-ROM).

APPENDIX B. EMAIL RECORDS[44]

Managing Records: a Quick Reference Guide For Electronic Mail And Other Documentary Materials

Email messages and other documentary materials are records when they:

- are made or received by an agency under Federal law or in connection with public business; and

- are preserved or are appropriate for preservation as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government, or because of the information value of the data they contain.

Principal Categories of Materials to Be Preserved:

- Records that document the formulation and execution of basic policies and decisions and the taking of necessary actions.

- Records that document important meetings.

- Records that facilitate action by Agency officials and their successors.

- Records that make possible a proper scrutiny by the Congress or by duly authorized agencies of the Government.

- Records that protect the financial, legal, and other rights of the Government and of persons directly affected by the Government's actions.

Email Messages and Other Documentary Materials That May Constitute Federal Records

[44] http://www.epa.gov/records/tools/quick.htm

- Materials providing key substantive comments on a draft action memorandum, if they add to a proper understanding of the formulation or execution of Agency action.

- Materials providing documentation of significant Agency decisions and commitments reached orally (person-to-person, by telecommunications, or in conference) and not otherwise documented in Agency files.

- Materials conveying information of value on important Agency activities, if they add to a proper understanding of Agency operations and responsibilities.

Email Messages and Other Documentary Materials That May Not Merit Extended Retention

- Materials, including email messages, that fall into one of the following categories:

    1) Non-record materials and personal papers, or
    2) Records requiring only short-term retention such as transitory documents; suspense, tracking, and control documents; facilitative documents; and selected disposable supporting materials.

- These types of documents are described in OIRM's "Procedures for Creating, Maintaining, and Disposing of Agency Records."

Managing Email Messages and Other Documentary Materials That Constitute Federal Records

- All Agency records must be included in a record keeping system, either manual or automated.

- Non-record materials, personal papers, and records not requiring long-tem retention should be filed separately.

- Some documents that qualify as Federal records only need to be retained for a relatively brief period of time.

Points to Remember about Email

- Agency email systems are for "official use" only by authorized personnel.
- Before deleting any email message, the author should determine whether it meets the legal definition of a record and, if so, preserve a copy of the message.
- Printed messages kept as a record should contain essential transmission and receipt data; if not, print the data or annotate the printed copy.
- Printed messages and essential transmission and receipt data should be filed with related files of the office.
- Messages that are not records may be deleted.
- When email is retained as a record, its retention period is governed by records disposition schedules.

APPENDIX C. DOD 5015.2 DIRECTIVE

Department of Defense

DIRECTIVE NUMBER 5015.2

April 11, 1997

ASD(C3I)

SUBJECT: DoD Records Management Program

References: (a) DoD Directive 5015.2, Records Management Program, March 22, 1991 (hereby canceled)

(b) Title 36, Code of Federal Regulations, Chapter XII, National Archives and Records Administration, Subchapter B, Records Management, current edition

(c) DoD 5025.1-M, DoD Directives System Procedures, August 1994, authorized by DoD Directive 5025.1, June 24, 1994

(d) Chapters 29, 31, 33, and 35 of title 44, United States Code

(e) DoD 5400.11-R, Department of Defense Privacy Program, August 1983, authorized by DoD Directive 5400.11, June 9, 1982

(f) DoD Directive 5100.3, Support of the Headquarters of Unified, Specified, and Subordinate Joint Commands, November 1, 1988

1. REISSUANCE AND PURPOSE

This Directive:

1.1. Reissues reference (a) establishing responsibility for the DoD Records Management Program, in accordance with reference (b), under the Assistant Secretary of Defense for Command, Control, Communications, and Intelligence (ASD(C3I));

1.2. Updates policy and responsibilities for life-cycle management (creation, maintenance and use, and disposition) of information as records in all media, including electronic; and

1.3. Authorizes the publication of implementing and procedural guidance on the management of records in the Department of Defense, consistent with references (b), (c), and (d).

2. APPLICABILITY

This Directive applies to the Office of the Secretary of Defense (OSD), the Military Departments, the Chairman of the Joint Chiefs of Staff, the Combatant Commands, the Inspector General of the Department of Defense, the Defense Agencies, and the DoD Field Activities (hereafter referred to collectively as the DoD Components).

3. DEFINITIONS

Terms used in this Directive are defined in enclosure 1.

4. POLICY

It is DoD policy to:

4.1. Create, maintain, and preserve information as records, in any media, that document the transaction of business and mission in wartime and peacetime to provide evidence of DoD Component organization, functions, policies, procedures, decisions, and activities as provided in Chapter XII of 36 CFR and Chapters 29, 31, 33, and 35 of 44 U.S.C. (references (b) and (d));

4.2. Manage records effectively and efficiently in compliance with this Directive and references (b) and (d), while protecting the legal and financial rights and interests of the Government and of persons affected by the Government's activities; and

4.3. Manage all records in any media used for creation or storage, in accordance with approved records schedules.

5. RESPONSIBILITIES

5.1. The Assistant Secretary of Defense for Command, Control, Communications, and Intelligence shall:

5.1.1. Issue guidance to implement this Directive and references (b) and (d), and address the following:

5.1.1.1. Policy necessary to establish, manage and maintain an active and continuing DoD Records Management Program under references (b) and (d); and

5.1.1.2. Procedures applicable to the creation, maintenance, use, preservation, and disposal of all records, in any storage medium, in compliance with references (b) and (d).

5.1.2. Improve and re-engineer DoD records management to enable OSD Principal Staff Assistants and the Chairman of the Joint Chiefs of Staff to manage information in records in their functional areas more effectively and efficiently.

5.1.3. Require the Defense Information Systems Agency to:

5.1.3.1. Establish and maintain a capability to test and evaluate automated records management information systems against legal, Agency-wide, and user requirements;

5.1.3.2. Establish and maintain a test and evaluation program for certifying automated records management information systems that meet the standard functional and automated system requirements for records management;

5.1.3.3. Review and coordinate all recommendations for changes to the DoD design criteria standard for records management functional baseline requirements, before approval by the Deputy Assistant Secretary of Defense (Command, Control and Communications); and

5.1.3.4. Establish and maintain a register of automated records management products that have been certified as meeting the standard functional and automated system requirements. Ready access to this register shall be provided to all DoD records management personnel.

5.2. The OSD Principal Staff Assistants and the Chairman of the Joint Chiefs of Staff shall:

5.2.1. Determine commonality of information in functional records management processes across the DoD Components to ensure information is available to support the warfighter.

5.2.2. Simplify and streamline records management within the Department of Defense by ensuring application of the principles and policies in section 4., above.

5.2.3. Evaluate, improve, implement, and execute DoD records management policies and procedures to ensure that functional management, control, oversight, and leadership are demonstrated during the life-cycle management of DoD records.

5.3. The Head of each DoD Component shall:

5.3.1. Establish and maintain the DoD Records Management Program at an organizational level of sufficient authority to ensure that the objectives and policies of this Directive and Chapters 29, 31, 33, and 35 of 44 U.S.C. (reference (d)) are efficiently and effectively implemented; and designate an individual to administer the DoD Records Management Program.

5.3.2. Apply standards, procedures, and techniques designed to improve the management of records, ensuring that records are:

5.3.2.1. Created, maintained, and preserved to document the organization, functions, policies, decisions, procedures, and essential operational, logistical, and support transactions of the Department of Defense as provided in 36 CFR XII and 44 U.S.C. 29, 31, 33, and 35 (references (b) and (d)) and DoD implementing Instructions and Publications; and

5.3.2.2. Created, maintained, and preserved to provide the information necessary to protect the legal and financial rights of the Government and of persons directly affected by DoD activities.

5.3.3. Use the most economical, efficient, and reliable means for creation, retrieval, maintenance, preservation, and disposition of records in any media.

5.3.4. Improve the management, maintenance, and security of records in coordination with OSD Principal Staff Assistants and the Chairman of the Joint Chiefs of Staff.

5.3.5. Apply DoD records management functional and system requirements to all electronic records management systems.

5.3.6. Incorporate records management requirements into automated information systems development and redesign.

5.3.7. Ensure proper training of all personnel that create and use records to ensure compliance with this Directive and references (b) and (d).

5.3.8. Advise all employees at least annually:

5.3.8.1. Of their responsibility to create and maintain records;

5.3.8.2. How to identify records and distinguish them from non-record materials;

5.3.8.3. Not to remove records from Government custody or destroy them, except as required or allowed under authorized record schedules;

5.3.8.4. How to inform appropriate officials of any actual, impending, or threatened unlawful removal, alteration, or destruction of Federal records; and

5.3.8.5. To identify personal papers and maintain them separately from organizational records, in compliance with reference (b).

5.3.9. Ensure prompt retirement or disposal of temporary records and the timely transfer of permanently valuable records under authorized record schedules.

5.3.10. Periodically evaluate the Components' compliance with the DoD Records Management Program and 36 CFR XII (reference (b)).

5.3.11. Advise the ASD(C3I) of records management issues that could have broad implications across the Department of Defense or between the Department of Defense and other Government Agencies, and fully cooperate with the ASD(C3I) in resolving these issues.

5.3.12. Safeguard all personal data within records, in accordance with DoD 5400.11-R (reference (e)).

5.4. The Secretaries of the Military Departments, in addition to the responsibilities in subsection 5.3., above, shall:

5.4.1. Provide guidance, training, assistance, and support to the Combatant Commands for the DoD Records Management Program in their area of responsibility, as outlined in DoD Directive 5100.3 (reference (f)).

5.4.2. Conduct periodic reviews of programs referred to in paragraph 5.4.1.,above; ensure all records are scheduled as provided in reference (b), and 44 U.S.C. 29, 31, 33, and 35 (reference (d)) and implementing DoD Instructions and Publications; and provide a written summary of the results of the reviews to the ASD(C3I). These reviews shall not be at the option of the Combatant Commanders or the other activities involved.

6. EFFECTIVE DATE

This Directive is effective immediately.

Enclosures - 1

1. Definitions

E1. ENCLOSURE 1

DEFINITIONS

E1.1.1. Agency. Includes the DoD Components and, any military, civilian or contractor personnel conducting operational, logistical, or support transactions anywhere within the DoD Components.

E1.1.2. Non-record. As defined in 36 CFR 1222.34 (reference (b)) Non-record materials are those Federally owned informational materials that do not meet the statutory definition of records (Section 3301 of reference (d)) or that have been excluded from coverage by the definition. Excluded materials are extra copies of documents kept only for reference, stocks of publications and processed documents, and library or museum materials intended solely for reference or exhibit.

E1.1.3. Record. As defined, in part, in Section 3301 of reference (d), Records include all books, papers, maps, photographs, machine-readable materials, and other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of the data in them. A record covers information in any medium, and includes operational logistics, support and other materials created or received by the DoD Components in training, contingency, and wartime operations as well as in all routine and peacetime business.

E1.1.4. Records Management. As defined in Section 2901 of reference (d), records management means the planning, controlling, directing, organizing, training, promoting, and other managerial activities involved with respect to records creation, records maintenance and use, and records disposition in order to achieve adequate and proper documentation of the policies and transactions of the Federal Government and effective and economical management of agency operations.

APPENDIX D. NARA, NRMP AND COURT RULINGS

BACKGROUND

This section provides background material on the subject of records management to familiarize the reader with the subject and terminology and to place into context the subsequent sections that address storage requirements. This section describes the governmental unit responsible for archive storage, the EPA program with records management oversight, and EPA's records and records sources.

U.S. National Archives and Records Administration (NARA) [45]

NARA is an independent Federal agency that helps preserve our nation's history by overseeing the management of all Federal records. The mission of NARA is to ensure ready access to the essential evidence that documents the rights of American citizens, the actions of Federal officials, and the national experience. NARA must make it easy for citizens to access this essential evidence regardless of the location of the documentation or of the people using it. NARA hopes that by providing electronic public access to more and more of their records and services, they will better meet the information needs of the American public.

The mission of the National Archives and Records Administration is rooted in legislation codified under Title 44 of the United States Code. Therein resides the authority of the Archivist of the United States, as head of the National Archives and Records Administration, to provide guidance and assistance to Federal officials on the management and disposition of records, to store records in centers from which agencies can retrieve them, and to take into archival facilities and Presidential libraries, for public use, records that are, in the language of Section 2107, "determined by the Archivist of the United States to have sufficient historical or other value to warrant their continued preservation by the United States Government."

As defined in Section 3301, these records are -

[45] Source for this sub-section: http://www.nara.gov/

> all books, papers, maps, photographs, machine readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of data in them.

Title 44, Section 3101 further specifies that -

> the head of each Federal agency shall make and preserve records containing adequate and proper documentation of the organization, functions, policies, decisions, procedures, and essential transactions of the agency and designed to furnish the information necessary to protect the legal and financial rights of the Government and of persons directly affected by the agency's activities.

NARA's four strategic goals are as follows:

> Essential evidence will be created, identified, appropriately scheduled, and managed for as long as needed.
>
> Essential evidence will be easy to access regardless of where it is or where users are for as long as needed.
>
> All records will be preserved in appropriate space for use as long as needed.

NARA's capabilities for making changes necessary to realize our vision will continuously expand.

U.S. EPA National Records Management Program (NRMP) [46]

[46] Source for this sub-section: http://www.epa.gov/records/about.htm

The NRMP [47] is responsible for providing leadership and direction for the Agency's national records management program. Among its responsibilities are developing an overall records management strategy producing the policy, procedures, and guidance necessary to implement that program, cooperating with other units in OIRM in developing policies and guidance on the application of technology to records management, coordinating the program within the Agency and with interested outside parties such as the National Archives and Records Administration, representing the Agency in interagency records management groups and assisting records programs across the Agency with advice and technical expertise, especially the Superfund and Dockets programs.

The role of the NRMP is to:

- develop and update internal Environmental Protection Agency policy and guidance,
- direct the EPA records disposition and retirement program,
- support communications among EPA records managers in the records management network,
- enhance the capability of the EPA Dockets and Records Centers,
- conduct briefings and training classes, and
- provide leadership in the application of technology to solve EPA records management problems.

The National program provides a wide range of services to EPA records managers and staff. It develops guidance products and training materials to meet the needs of the Agency's records management staff; responds to requests for technical assistance on all aspects of records management; operates the Docket and Records Management Networks; coordinates communications and networking among records and dockets managers Agency wide; promotes improved communications among records and dockets managers through meetings, site visits, and conferences; publishes INFO ACCESS a bimonthly newsletter and numerous other publications; and coordinates relations with the National Archives and Records Administration.

[47] NRMP, Mail Code 3408, 401 M Street, SW Washington, DC 20460, Phone: 202-260-5926, Fax: 202-401-5447, Email: nrmp@epa.gov

The NRMP staff maintains the official set of Agency records disposition schedules, which serve as a high level finding aid to all Agency records. The National program staff also maintains a complete set of Agency records management policy, guidance, and procedures, as well as a reference collection of records management related materials.

Court Rulings on Electronic Records

In 1972, [NARA] began authorizing disposal of electronic records through general schedules in very limited circumstances. In October, 1994, the Acting Archivist proposed revisions to GRS 20 to explicitly authorize all federal agencies to destroy agency records stored on word processing and electronic mail systems, as well as other electronic records, once the records have been printed in "hard copy" on paper or microform, and if the agency determines that it no longer needs the electronic version.[48]

On August 14, 1995, [NARA] issued the present version of GRS 20 which authorizes the disposal of electronic records in fifteen enumerated categories, including electronic records created by computer operators, programmers, analysts, systems administrators and government staff using office automation applications. Specifically, GRS 20, Item 13 provides that word processing files recorded on electronic media, "after they have been copied to an electronic record keeping system, paper, or microform for record keeping purposes," may be deleted from the word processing system "when no longer needed for updating or revision." GRS 20, Item 14, covering electronic mail records, provides that senders' and recipients' versions of electronic mail messages and attachments thereto may be deleted from the electronic mail system after they have been copied to an electronic record keeping system, paper or microform for record keeping purposes.[49]

[48] U.S. District Court, District of Columbia October 22, 1997 ruling. (http://www.nara.gov/records/grs20/opinion.html)

[49] Ibid

On October 22, 1997, United States District Judge Paul L. Friedman ruled in the case of Public Citizen, Inc., et Al. Vs. John Carlin, Archivist of The United States, et Al., that the plaintiffs' motion for Summary Judgment be granted and that the court declare that General Records Schedule 20 (GRS 20) is null and void. He further ordered that the court declare that defendant agencies executive office of the president, office of administration, and the United States trade representative may not destroy electronic records created, received or stored on electronic mail or word processing systems pursuant to general records schedule 20. In its ruling, the court cited the following as one reason for preserving the electronic version of paper records:

> "Finally and most importantly, electronic records often contain information that is not preserved in a print-out record or even in other computerized systems of records. For example, paper print-outs of computer spreadsheets only display the results of calculations made on the spreadsheet, while the actual electronic version of the spreadsheet will show the formula used to make the calculations."[50]

On August 6, 1999, an Appeals from the United States District Court for the District of Columbia reversed the October 22, 1997 ruling and upheld the GRS 20 policy promulgated by the archivist at NARA. NARA has issued the following statement[51]:

> In 1995, Archivist of the United States John Carlin, as head of the National Archives and Records Administration, issued a revision of a general records schedule, GRS 20, for the disposition of some Federal electronic records. GRS 20 authorized federal agencies to dispose of certain electronic copies of email, word-processing documents, and other computer-generated material if first saved by copying to a paper, microform, or electronic record keeping system. In response to a lawsuit opposing this action, a Federal district court judge ruled that portions of GRS 20 reached beyond the authority of the Archivist of the United States under the Federal Records Act, and declared

[50] http://www.cadc.uscourts.gov/common/opinions/199908/97-5356a.txt

[51] Statement by the National Archives and Records Administration on the U.S. Court of Appeals decision on General Records Schedule 20. (http://www.nara.gov/records/grs20/state806.html)

> GRS 20 null and void. In response to the Government's appeal, the U.S. Court of Appeals has now overturned the lower court decision.
>
> The National Archives and Records Administration is pleased that the court has upheld the legal authority of the Archivist and the propriety of his action. We now need to review the appeals court's decision before making any further comment. But NARA welcomes the opportunity provided by this decision to continue in an orderly way to develop practical, workable strategies and methods for managing and preserving records in the electronic age and ensuring ready access to them. NARA remains committed to working aggressively toward that goal.

Although the court agreed that maintaining email and word processing records digitally "has advantages over paper record keeping," it concluded that an agency legally can be allowed "to choose, based upon its own operational needs, whether to use electronic or paper record keeping systems."

Since the appeals court ruling is so recent, there has not been time for NARA or for EPA's NRMP to evaluate and possibly change the guidelines currently in place for archiving. This paper therefore is forced to base its analysis on current NRMP policies and procedures.

APPENDIX E. EXAMPLE RECORD SCHEDULES

Example No. 1 - EPA Records Schedule for the Files on Ocean Dumping

EPA SERIES NO. 383H U.S. EPA RECORDS CONTROL SCHEDULE

SERIES TITLE: Ocean Dumping Files

PROGRAM: International Activities

EPA SERIES NO: 383H

AGENCY FILE CODE: INTE 383

NARA SCHEDULE NO. NC1-412-78-8/11
(Use this number to retire records to the FRC)

APPLICABILITY: Headquarters

IDENTIFYING INFORMATION:
DESCRIPTION: Includes records relating to EPA participation in implementing the 1972 Ocean Dumping Treaty. Records relate to EPA chairmanship of the U.S. delegations to treaty group meetings, Agency monitoring of ocean dumping incidents, review and evaluation of ocean dumping regulations, and other related matters. Records consist of position papers, conference reports, correspondence, and other related records.

ARRANGEMENT:

TYPE OF RECORDS: SPECIFIC RESTRICTIONS:
Case files

MEDIUM: VITAL RECORD:

Paper, electronic

FUNCTIONS SUPPORTED:
Program operations

SPECIFIC LEGAL REQUIREMENTS:

DISPOSITION INFORMATION:

FINAL DISPOSITION:	TRANSFER TO FRC PERMITTED:
Disposable	Yes

FILE BREAK INSTRUCTIONS: Break file at the end of each year.

DISPOSITION INSTRUCTIONS: Keep in office 3 years after file break, then retire to FRC. Destroy when 10 years old.

APPLICATION GUIDANCE:
REASONS FOR DISPOSITION: Disposition previously approved by the National Archives.

AGENCY-WIDE GUIDANCE:

PROGRAM OFFICE GUIDANCE/DESCRIPTIVE INFORMATION:

CUSTODIAL INFORMATION:

CONTROLLING UNIT:	CONTACT POINT:
Name:	Name:
Location:	Mail Code:
Inclusive Dates:	Telephone:
Volume on Hand (Feet):	Office:

Annual Accumulation: (feet or inches)

Room:

CONTROL INFORMATION:
RELATED ITEMS:

PREVIOUSLY APPROVED BY
NARA SCHEDULE NOS:

Approval Date EPA	Approval Date NARA	Entry Date	Last Modified
3/3/78	7/8/78	12/8/92	2/25/98

Example No.2 - EPA Records Schedule for Personnel Correspondence and Forms

EPA SERIES NO. 561 U.S. EPA RECORDS SCHEDULE

SERIES TITLE: Personnel Correspondence and Forms Files

PROGRAM: Personnel

EPA SERIES NO: 561

AGENCY FILE CODE: PERS 561

NARA DISPOSAL AUTHORITY: General Records Schedule 1/17 (Items a-c)
General Records Schedule 1/43 (Item d)

APPLICABILITY: Agency-wide

IDENTIFYING INFORMATION:
DESCRIPTION: Contains operating personnel office records relating to individual employees not maintained in the OPFs and not provided for elsewhere:

a. Correspondence and forms relating to pending personnel actions.

b. Retention registers and related records:

(1). Registers and related records used to effect reduction-in-force actions.

(2). Registers from which no reduction-in-force actions have been taken and related records.

c. All other correspondence and forms.

d. Electronic copies of records that are created on electronic mail and word processing systems.

(1) Copies that have no further administrative value after the recordkeeping copy is made. Includes copies maintained by individuals in personal files, personal electronic mail directories, or other personal directories on hard disk or network drives, and copies on shared network drives that are used only to produce the recordkeeping copy.

(2) Copies used for dissemination, revision, or updating that are maintained in addition to the recordkeeping copy.

ARRANGEMENT:

TYPE OF RECORDS: SPECIFIC RESTRICTIONS:
Case files

MEDIUM: VITAL RECORD:
Paper, electronic

FUNCTIONS SUPPORTED:
Personnel administration

SPECIFIC LEGAL REQUIREMENTS:

DISPOSITION INFORMATION:
FINAL DISPOSITION: TRANSFER TO FRC PERMITTED:
Disposable No

FILE BREAK INSTRUCTIONS: See Disposition Instructions.

DISPOSITION INSTRUCTIONS:
a. Destroy when action is complete.

b(1). Destroy when 2 years old.

(2). Destroy when superseded or obsolete.

c. Destroy when 6 months old.

d(1). Destroy/delete within 180 days after the recordkeeping copy has been produced.

(2). Destroy/delete when dissemination, revision, or updating is completed.

APPLICATION GUIDANCE:
REASONS FOR DISPOSITION: Conforms to NARA's General Records Schedule 1, items 17 and 43.

AGENCY-WIDE GUIDANCE: As required by 36 CFR 1228.58, records with special restrictions such as confidential business information (CBI) or those exempted from disclosure by statute such as the Privacy Act must be shredded or otherwise definitively destroyed with witnessed disposal for records destroyed by contractors.

PROGRAM OFFICE GUIDANCE/DESCRIPTIVE INFORMATION:

CUSTODIAL INFORMATION:

CONTROLLING UNIT:	CONTACT POINT:
Name: Personnel Office	Name:
Location:	Mail Code:
Inclusive Dates:	Telephone:
Volume on Hand (Feet):	Office:

Annual Accumulation: (feet or inches)

Room:

CONTROL INFORMATION:
RELATED ITEMS:

PREVIOUSLY APPROVED BY
NARA SCHEDULE NOS: NC-412-75-4/8, NC1-412-85-28/11

Approval Date EPA	Approval Date NARA	Entry Date	Last Modified
NA	12/98	7/30/93	7/8/99